Cluster Analysis: Algorithms, Hazards and Small Area Relative Survival

September 2008

Ceri White

Faculty of Advanced Technology
University of Glamorgan
Treforest, Pontypridd

A submission presented in partial fulfilment of the requirements of the University of Glamorgan/Prifysgol Morgannwg for the degree of Doctor of Philosophy

TABLE OF CONTENTS

Acknowledgements

I would like to express my thanks to my supervisors, Prof. J.A. Ware, Dr. G. Higgs, Prof. F. Dunstan and Dr. J.A. Steward for their continued advice and support throughout the duration of the research.

I would also like to thank the Welsh Cancer Intelligence and Surveillance Unit in Cardiff for allowing me to carry out this research on a part-time basis at the University of Glamorgan and for the data that was required for analysis for the research.

Finally, I would like to thank my family for their continued love and support.

Certificate of Research

This is to certify that, except where specific reference is made, the work presented in this thesis is the result of the investigation undertaken by the candidate.

Candidate .....

Director of Studies .....

Certificate of Research

This is to certify that neither this thesis nor any part of it has been presented or is being currently submitted in candidature for any other degree other than the degree of Philosophy of the University of Glamorgan.

Candidate .....

Abstract

Cluster Analysis: Algorithms, Hazards and Small Area Relative Survival

This thesis presents research that has demonstrated the use of clustering algorithms in the analysis of datasets routinely collected by cancer registries. This involved a review of existing algorithms and their application in studies of spatial and temporal variations in cancer rates. As a result of continuing public and scientific concern there has been an increase in the numbers of cancer related enquiries in recent years that has helped to raise the profile of the work of cancer registries. There are no official guidelines on the approach to be taken in such studies in relation to cluster analysis. In this study, a variety of cluster algorithms were applied to leukaemia data collected by the Welsh Cancer Intelligence and Surveillance Unit in order to propose an approach that could be adopted in future investigations of cancer incidence in Wales. For example, different methodologies have been employed to determine if an excess risk occurs near hazardous sources and one of the studies in the portfolio compares the results of using three methods to determine if an increased risk of cancer occurs in the vicinity of landfill sites and electric power lines. This uses new digital products that permit a more detailed estimation of the population at risk and permit a sensitivity analysis of the results of such investigations. In the third portfolio, analysis of relative survival at small area level has been made possible using a new level of geographical resolution that has recently been released in the United Kingdom. This study shows the benefits of using this new level of geography for small area studies of cancer survival where there are generally small numbers of deaths per spatial unit. It is anticipated that together these research studies will be of wider benefit to other registries in the UK charged with investigating spatial and temporal variations in cancer rates.

1. CRITICAL OVERVIEW

Introduction

Disease incidence varies spatially between areas. This can give rise to clues regarding the aetiology and risk factors for particular diseases. Hazardous sources placed in particular areas may have caused the incidence to increase. There have been many studies examining whether increased risks exist around potential hazardous sources such as those by Dolk et al. (1998), Vrijheid et al. (2002) and Elliott et al. (2001). Other studies have examined the spatial patterns such as those by Clayton and Kaldor (1987), Devine et al. (1994) and Waller et al. (1997). Various problems arise from such studies because particular covariates such as deprivation are not taken into account. Thus, particular methods and techniques are required in order to assess these. One particular area of concern is the risk of cancer in relation to hazardous sources.

The work of cancer registries in the United Kingdom increasingly involves cluster analysis, driven by public demands for information and assurances regarding the impact of hazards such as radiation on rates of cancer. There are eight regional English cancer registries and three national cancer registries for Northern Ireland, Scotland and Wales. The Welsh Cancer Intelligence and Surveillance Unit (WCISU) was established in 1997 when the registration of cancers was transferred to Velindre NHS Trust. All cancer registrations along with death information on cancer patients are collected from various sources such as hospital episodes, pathology, death certificates for the resident population of Wales from 1974 to the present day. Considerable work has been conducted regarding past cluster enquiries, in particular an alleged "cluster" of childhood cancers living in the Chepstow area, within 10km of Oldbury nuclear power station (WCISU, 2002). The WCISU actively respond to various concerns regarding alleged cancer clusters, putative hazards and also small area variations in incidence, mortality, survival and outcome. The research documented in this thesis aims to investigate these types of queries using Geographical Information Systems (GIS) and statistical software and to make

recommendations regarding the future analysis of clusters that may have direct relevance for the work of other cancer registries both in the UK and beyond.

The WCISU is a member of the United Kingdom Association of Cancer Registries (UKACR) which is actively involved in the area of cancer clustering and geographical mapping. At present, cancer registries adopt different methodologies with regard to the analysis of cancer clusters. A questionnaire was sent to a statistician or information analyst at each UK Cancer Registry during July 2003 in order to gauge the current state of play with regards to potential clusters of cancer. Findings presented in *Appendix A* show that at that time many of the UK cancer registries did not use clustering algorithms. As a result of continuing public and scientific concern there has been an increase in the numbers of cancer related enquiries in recent years that has helped to raise the profile of the work of cancer registries. However, there are no official guidelines on the approach to be taken in such studies in relation to cluster analysis. The aim of these studies is to compare the use of such techniques in relation to the types of enquiries that are addressed to registries in the UK.

Various issues have to be addressed such as the data that are available to use. There are various geography levels in the UK that could be used. Additionally, case data can be individual to the person or aggregated to a specific geography level. Aggregated data will inevitably produce different results compared with individual level data. The type of data to use depends on the requirements of the technique. However, problems tend to arise when the corresponding population figures are needed for analysis. Population figures are only available for specific geography levels; hence the case data and geographical unit used will depend on this factor. Individual level data tend to be used for case and control studies where population figures are not required, hence the increased accuracy at small area level. It is generally found that case and population at risk data require aggregated information since population figures are not available at individual level.

There are various methods used in the literature to investigate clusters and the definition of the exposed area. Different techniques will adopt different methods with respect to the determination of the population at risk in a cluster. For example, one method may identify all possible clusters for all population sizes within a specific set of centroids for a maximum defined population figure whereas another method may only detect clusters for that specific defined population size. Thus, the selection of the technique to use is an important factor as resulting clusters may, and more often may not, vary from one method to another.

A majority of methods examine the risk of incidence of disease but risk of outcome is also an important factor, in particular survival. Survival is calculated via the diagnosis date and death date of a patient. Analysis of survival at small area level will identify areas where survival is much better (or worse) compared with other neighbouring areas that would not have been identified otherwise. These areas can then be further explored as to possible reasons why this could be the case.

The research documented in this thesis consists of three separate, but intrinsically linked themes regarding cluster analysis. The first theme examines a variety of clustering algorithms contained in the software package ClusterSeer V2.2.4. The second theme examines population based techniques used to determine whether an increased risk of cancer is associated with potential hazardous sources using digital products that have recently become available in the UK. The third theme investigates relative survival at small area level using recently defined Middle Super Output Areas. The first area of research involved a comparison of existing clustering algorithms to determine the most effective technique. Effectiveness in this instance was defined in terms of advantages, disadvantages, comparability of results between other algorithms, multiple runs and ability to consistently locate the same actual clusters when parameters were varied in the analysis of cancer clusters. This was not simply an exercise of applying a range of techniques to datasets but to find the most effective algorithm to use for cluster enquiries to the WCISU (WCISU 2002, 2005) which may mirror those to other cancer registries in the UK. Cancer datasets from the WCISU were used to compare the findings from

applying each of the algorithms. The most effective algorithm identified in the initial theme is used in theme two which examines the population potentially at risk in the 'exposed' area of interest around landfill sites and power lines. These techniques were also used in theme three which involved an investigation of spatial variation in small area relative survival rates in Wales. The aim from the outset was not to develop new clustering algorithms; rather the research was confined to the use of existing algorithms in order to gauge their use in spatial and temporal analysis of cancer rates.

There were a number of disciplines encapsulated in this research including computer science, geographical science and statistical science. However, other areas such as epidemiology, Public Health, demography and cancer registration were heavily utilised. The research was divided into three areas of investigation, all of which had the common theme of "cancer clustering" and spatial variation of cancer in Wales.

Theme One

Aims and objectives (theme one)

The main aims and objectives of the first theme were as follows.

- To examine the various clustering algorithms included in the software package ClusterSeer V2.2.4 and SaTScan V5.1.3 using real data sets and simulated datasets of case and population at risk data or case and control data.
- Compare results of the clustering algorithms (for those which could be compared to each other) to determine the most effective algorithm when dealing with cancer cluster enquiries at cancer registries in the UK.

It was reported by Paul Elliott at a Spatial Epidemiology Conference in Spring 2006 that past studies in relation to investigations of cancer incidence and hazardous sources tend to ignore comparisons between different algorithms. It was concluded by Paul Elliott at the conference that a large number of clustering algorithms were required in order to

facilitate a comparison of the findings of different algorithms to determine which was the 'best' algorithm to use in particular studies in terms of effectiveness. The first theme of this research concentrated on an investigation of those clustering algorithms that could potentially be used to detect various types of clustering using a leukaemia dataset in Wales for a twenty year period 1982-2001. The results from the algorithms were compared to determine how well each test performed with regard to other existing algorithms (for those that could be compared with each other). Additionally, simulated datasets were used to identify clusters to determine if such algorithms identified the clusters that had been deliberately included into such datasets. The "most appropriate" clustering algorithm, in terms of practicality, suitability and strengths over other algorithms was found and applied to theme two and theme three of this portfolio.

Methodology (theme one)

There are various types of clustering. For example, a person may be interested in whether a cluster exists around a particular point source or whether an increased risk exists in a particular area. Algorithms to identify various types of clustering were applied to the datasets using the software package ClusterSeer V2.2.4. Each algorithm was examined using a leukaemia dataset for the twenty year period 1982-2001 for Welsh residents in order to identify clustering in the dataset. Additionally, the software SaTSCan V5.1.3 was used due to a problem with memory when examining this algorithm in ClusterSeer V2.2.4. Those algorithms that required the same type of data, case and population at risk data or case and control data at the same aggregated or point level were compared with each other, in terms of parameter changes, to determine which, if any, produced consistent results in the identification of clusters in Wales. Eight simulated datasets that contained artificial clusters were examined using clustering algorithms for those algorithms that could be compared with each other regarding point or areal data in order to determine how well the algorithms performed at locating these clusters.

Results and conclusions (theme one)

Some of the methods examined (Moran's I statistic, Oden's I Pop statistic, Besag and Newell's global method and Cuzick and Edwards' method) did not identify specific clusters in the datasets; they examined whether the dataset in question showed any evidence of clustering. This is slightly different with regards to specific locations of clustering. One such method, Moran's I statistic used rate data only (2.4.1), thus the population differences between small areas in Wales were not accounted for; a high rate may have been based on just two cases – thus a decrease of one case would have halved the resulting rate. Some clustering algorithms, in particular Besag and Newell's method were disregarded due to their lack of consistency when changing the parameter for the maximum cluster size, for example increasing the maximum cluster size in any cluster by one observed case changed the resulting conclusion from a highly significant result to a highly non-significant result for one clustering algorithm (2.4.3). It was noted that resulting p-values using two of the clustering algorithms (Score test of Lawson and Waller and Bithell's linear risk score test) to determine if an increased risk existed around a particular focus changed dramatically depending on the distance used in the dataset (2.62, 2.63). That is, if the two datasets were created from the original dataset for two different radii from the focus (e.g. 5km and 10km). The corresponding p-values obtained could vary depending on the distribution of cases in one dataset and not the other since the algorithms use each dataset to calculate the background rate to determine if clustering exists around the focus. Caution is advised when values are input by the user for some algorithms in ClusterSeer V2.2.4, especially if looking at a small area of interest with a small population since resulting p-values can change dramatically. When using the simulated datasets (2.91, 2.92), the spatial scan statistic by Kulldorff was able to locate the clusters, irrespective of the maximum population size used. It should be noted that when two clusters overlapped each other, this method detected one larger cluster as opposed to two smaller clusters. However, no other clustering algorithm was able to detect the two smaller clusters that overlapped.

One of the main limitations in ClusterSeer V2.2.4 was the lack of confounding available for many of the clustering algorithms; a problem with the software. Results have been shown to change if particular factors such as age and deprivation are taken into account. Only two of the methods, Diggle's method and Cuzick and Edwards' method, those which used case and control data in ClusterSeer V2.2.4, were able to take confounding into account. SaTScan V5.1.3 is able to take such factors into account. Thus, for this area of research, analysis of the datasets did not account for confounding. This is explored further in theme two and theme three. It was concluded that the spatial scan statistic by Kulldorff that SaTScan is based on (2.53, 2.61, 2.71), was the most effective at detecting clusters in Wales compared to other algorithms when using case and population at risk data. Not only does it have the advantage in that it can detect various types of clustering such as around a focal point or searching for a small cluster in a large area, it could identify the same cluster when varying the maximum population size that was contained in the cluster as long as the maximum population size was larger than the population at risk in the actual cluster. However, this method did have a problem in identifying clusters when clusters overlapped but no other algorithm examined could overcome this issue (2.9.2.2). Another clustering algorithm, Anselin's local Moran test proved useful in that it could detect areas that were similar or dissimilar to other neighbouring areas in terms of the observed number of cases (2.5.4). This algorithm should also be used to identify these outliers. These outliers can be investigated further, for example in an instance where one area may contain a much higher number of incident cases compared with neighbouring areas (as would possibly be the case if there was a nursing home situated in the ward). Another advantage that the spatial scan statistic had over other algorithms was that case and control data could be examined and the user is able to determine whether the resulting clusters should overlap or not. An important conclusion for this project was that the algorithm should be recommended for future analysis at the WCISU and may be of interest for other cancer registries in the United Kingdom when dealing with such cluster enquiries.

Theme Two

Aims and objectives (theme two)

There were several objectives in this theme. These were evaluated by analysing datasets of cancer from the WCISU in relation to point hazardous sources and linear hazardous sources.

- To describe in detail the techniques used to aggregate population at risk for any exposed area.
- To compare the results of these methods to determine the extent at which resulting conclusions agree or disagree. Extrapolation of population, definition of the "unexposed" population at risk and choice of geographical unit were explored to determine if these influenced results.
- To determine whether an increased risk of cancer existed around landfill sites and electric power lines in Wales.
- To investigate the effect of latency periods, the time from exposure to disease, on most likely clusters in Wales.
- To examine the comparability of results of all techniques examined with the spatial scan statistic by Kulldorff, the most effective algorithm from theme one when taking the latency period into account.

The second area of study concentrated on the methodology of selecting a reference population to determine if significant increased risks existed around hazardous sources. The aggregation of the data to be analysed and the level of geography used affects the results depending on the technique used. These were explored in detail using two traditional techniques (intersection method and centroid method) and a more recent technique to estimate the population at risk in an exposed area using postcoded data. Resulting conclusions regarding significance were compared. The more recent technique was developed by researchers at the Small Area Health Statistics Unit (SAHSU) at Imperial College, London to identify the population "at risk" within a specific area of interest using postcode data (Briggs et al., 2001). Previous studies have tended to use just

one method and make conclusions regarding increased or decreased risks in the exposed areas based on this method. This study involved an investigation of the choice of geographical unit used along with the extrapolation of population and the "unexposed" population at risk used to calculate expected counts using the three population based techniques. The results were examined and compared between each method to determine the influence of these factors. Cancer registries usually receive a cancer cluster enquiry regarding one landfill site so additionally, the three techniques were implemented for this scenario and comparisons made. This theme also included an exploration of these methods using linear hazardous sources as well as point sources. The latency period, the time it takes for the onset of disease, was also taken into account to determine the significance of these clusters. The "most appropriate" clustering algorithm from the first theme was used to identify clusters in Wales in order to compare the significance of the clusters when each technique was used when taking the latency period into account.

Methodology (theme two)

The problem that investigators are posed with is defining a region of exposure at a particular geographic level at which data are available. To explore this further, the three population estimation techniques used two geography levels available in Wales to determine if resulting conclusions differed. The observed number of cases, expected number of cases, standardised incidence ratios (the ratio of those observed to those expected) were calculated and compared in different scenarios for each of the three techniques in order to assess the validity of the results obtained. This was investigated around hazardous point sources and hazardous linear sources in Wales using cancer datasets from the WCISU. Various assumptions regarding the population at risk were made in order to use the techniques.

Other factors were introduced such as adjustment of population to take into account the population at risk for inter census years. The operation dates of the hazardous sources and calculation of expected values using background rates were explored to compare the

resulting conclusions with previous techniques. The results were adjusted using latency periods that past studies had adopted to determine if results were consistent.

Results and conclusions (theme two)

It was found that two of the methods, the centroid method and postcode method produced similar resultant standardised incidence ratios with a slightly different number of observed and expected cases (table 3.7) around hazardous point sources. The intersection method was very poor due to the aggregation method used which resulted in a much higher number of observed cases and population at risk in the exposed area compared with the other two methods. This method had the disadvantage of diluting the effect of an increased risk, if one did exist. Similarly, the centroid method included geographical units (and hence people) outside the buffer and some geographical units (and people) were excluded that were contained within the buffer, hence, this method too had the effect of diluting the effect of an increased risk, if one did exist or possibly increasing the risk if a true decreased risk existed. The choice of geographical unit used affected the results of the two traditional methods (intersection method and centroid method) but did not affect the results of the more recent postcode method (table 3.7). Extrapolation of population was used to adjust the results to determine whether this factor affected the conclusions (table 3.8). The number of expected cases and resulting standardised incidence ratios only slightly changed when this was taken into consideration as did the results when using all Wales or those cases and population at risk over a particular distance from the point sources as the background rate (table 3.10). Taking into account the operation times of the hazardous point sources only slightly changed the resulting standardised incidence ratios for the cancers examined (table 3.11). The latency period, the time from exposure to disease, affected the results in terms of observed numbers and expected numbers when examining clusters that the most effective algorithm located around the point source (table 3.14, table 3.16). Care should also be taken when working with small numbers since an increase of just one or two cases can dramatically affect the significance of the results. Hence, if a borderline significant result was obtained before these factors were taken into account (not seen in the results in this thesis), then the

resulting conclusion could well change. There appeared to be a difference in results regarding significance when analysing the risk around just one point source (table 3.11), the case that most cancer registries will have to investigate and hence the need for an accurate definition of the population at risk in the exposed area. When examining a hazardous linear source, the number of observed cases in the analysis appeared to be very similar for the centroid method and postcode method (table 3.18). However, on further investigation, it was found that approximately only a small proportion of the observed cases in the exposed area using one of those techniques were contained in the exposed area using the other population estimation technique due to the selection of geographical units within the distance studied (table 3.19), thus the importance of correctly identifying the population at risk in the exposed area. It was concluded in the absence of true numbers of population living in a potentially exposed area that the recent postcode method adopted by SAHSU for estimating the population at risk should be used in the analysis of hazardous sources at cancer registries in the UK. This technique enabled an improvement in aggregating population at risk within a specific area of interest using new digital products that have become recently available. This technique involved the use of an estimate to include only those cases inside the specific area of interest and aggregated the population, with underlying assumptions.

This area of research was not an epidemiological study but a methodological study in terms of comparing existing methods used to identify if an increased risk exists. It is suggested that such tests can be used as an exploratory analysis prior to a more detailed analysis based on factors such as home electrical measurements and the wind and speed direction which could influence the exposed area of interest. A literature review of past studies is warranted to determine the latency period of the cancer, as well as the operation period of particular hazardous sources examined and should also be taken into account in the analysis.

Theme Three

Aims and objectives (theme three)

The following aims and objectives were set for this theme:

- To determine whether relative survival rates of particular cancers differed between regions in Wales.
- To examine relative survival rates of these cancers at a smaller geographical unit to determine any spatial patterns that may exist.
- To investigate areas of high and low relative survival rates via the most effective algorithm from theme one.
- To investigate areas of high and low relative survival rates via various smoothing methods.

Very little work has been carried out on small area survival analysis in the world. The third study aimed to facilitate the identification of spatial patterns of survival rates with the aim of informing remedial action that will hopefully improve survival rates in these areas. Mapping aids the user as a visualisation tool in order to identify areas of high or low rates throughout a study region. Bayesian methods, a type of smoothing method, consider prior information on the variability of disease rates in the overall map, in addition to the observed events in the area. In general, smoothing methods tend to "average out" neighbouring areas to overcome the problem of few events per geographical unit. The third study examined the relative survival rates around neighbouring geographical units. Relative survival, the ratio of observed survival to expected survival for two cancer sites, namely female breast cancer and colorectal cancer in Wales was examined at small area level, specifically at Middle Super Output Area (MSOA) level, a geography level defined by output areas from the 2001 UK Census. This development has enabled the opportunity of robust estimates when calculating relative survival rates at small area level. Age, sex and deprivation were taken into consideration. In previous studies, problems have arisen with regard to survival and small area analysis due to the small number of events (deaths). The introduction of

MSOAs (413 in Wales) enabled survival estimates to be calculated at this level to enable more statistical reliability. Small area survival estimates proved very difficult in the past due to the very small populations at risk at ward and other existing geographical units. Due to varying survival estimates between neighbouring areas, smoothing was used to identify local areas of high and low survival. This theme adopted a Bayesian approach - a technique that "averages out" the survival estimates between neighbouring geographical units. The most effective algorithm from theme one was used to identify high and low survival rate clusters and compared with the smoothed analysis to determine if both methods showed similar high and low rate clusters. Additionally, the algorithm that identified outliers in the first theme was used to determine if any spatial pattern was evident in the datasets. Mammography screening tends to diagnose breast cancer tumours earlier than would have been expected. Thus, the tumour is less advanced than would have been without screening and survival improves by at least the time until it would have been diagnosed (Antinnen et al., 2006, Tabar et al., 2003). This factor was included in the smoothed model to compare with the pre-screening model.

Methodology (theme three)

Relative survival rates for female breast cancer and colorectal cancer were calculated for Wales, Local Health Boards in Wales and MSOAs in Wales using an algorithm in the statistical software package STATA (Esteve et al., 1990). Life tables, a summary table of all cause mortality, were created which adjusted for age, sex and deprivation since available life tables do not adjust for deprivation. The relative survival rates between neighbouring areas were examined via the use of GIS. Three different time periods were examined for female breast cancer. These three periods were also examined for colorectal cancer as well as for males and females individually. It was difficult to observe any spatial pattern between neighbouring areas using the maps of relative survival by MSOA due to the varying rates at small area. Hence, Bayesian methods were explored to determine if any spatial pattern existed. The most effective algorithm from the first theme was used to determine local clusters of high and low survival rates to compare with the areas of high and low survival rates using smoothing. The percentage

of women that were screened in each Local Health board was obtained from Breast Test Wales, and was used in the smoothing model to determine if screening affected the smoothing of the survival rates.

Results and conclusions (theme three)

Examination of relative survival rates between neighbouring regions in Wales (figure 4.3) showed varying rates which warranted further investigation at a lower level of geography. Middle Super Output Areas were used since at least ten deaths are required in an individual calculation for a reliable survival estimate and lower level geographical units do not satisfy this criterion (table 4.5). There were still a small number of MSOAs that did not satisfy this criterion when examining the individual ten year periods, especially for female breast cancer so caution is advised when interpreting these results for the few MSOA that do not satisfy the criterion. Due to the resolution of geographical units, it was unclear whether nearest neighbours of MSOAs had similar or dissimilar relative survival rates (figures 4.4 to 4.6). The smoothing of these relative survival rates showed survival rates very close to the mean for the whole of Wales for female breast cancer for the three periods studied (figure 4.7). It appeared that the smoothing model had "over smoothed" the data and was therefore difficult to observe localised areas (or clusters) of high or low survival. Colorectal cancer however, did show neighbouring areas in Wales having higher (or lower) relative survival rates compared with the rest of Wales (figures 4.8 and 4.9). Advantages of identifying areas of high and low survival in this way are that these areas can be investigated further for possible reasons as to why this was the case; for example to examine the potential role of distance to the hospital of treatment. The most effective algorithm from theme one was used to locate clusters of low and high survival rates in Wales for female breast cancer and colorectal cancer for the entire twenty year period 1982-2001 (figure 4.13 and 4.14). For female breast cancer, the clusters obtained could not be compared with the smoothed data due to the over smoothing of the dataset using Bayesian methods. However, for colorectal cancer, it did appear that there was some consistency between the clusters that the spatial scan statistic

detected and the spatial pattern of high and low relative survival rates using the smoothing method. Another clustering algorithm was used from theme one to determine if any spatial patterns existed and showed evidence of such patterns in all datasets. The largest of these spatial variations was seen in colorectal cancer for a ten year period. The corresponding analysis for female breast cancer also showed that neighbouring areas had similar rates (table 4.8). Displaying all primary and secondary clusters that the most effective algorithm detected on a map, identified a large number of high rate clusters near to each other for female breast cancer, which was not seen for colorectal cancer (figure 4.15). This could be the reason as to why the smoothing model for female breast cancer over smoothed the data. The smoothed model for female breast cancer did not appear to markedly change when taking into account breast screening (figure 4.16). The calculation of relative survival rates at small area level has identified local areas of clustering that would not otherwise have been determined. These areas can be further explored as to reasons for variation in survival rates which could involve investigating the incidence and mortality rates in these areas, analysing travel times to screening or treatment programme or by examining detailed histories of patients.

Limited analysis was conducted with regards to other factors that could affect relative survival due to the lack of detail available on cancer patients in the relevant study period. Case note validation could have overcome this but would have involved an enormous amount of work due to the number of cases of female breast cancer and colorectal cancer in the dataset. From the mid 1990s, the WCISU have collected information regarding treatment and stage which could also affect the relative survival rates. Additionally, only the twenty year period was analysed using the spatial scan statistic for both cancers. The ten year periods may also have showed comparisons between the smoothing model and the detected high and low rate clusters.

Bringing the three themes together

As shown in this overview of the three themes, there is a common theme of cancer clustering and the spatial variation of cancer at small area level that is central to each of the themes. The first area of research guides the analysis conducted in the second and third themes in relation to the use of specific clustering algorithms. All three themes analyse data using low level geographical units, namely wards for theme one, wards and enumeration districts for theme two and MSOAs for theme three. MSOAs were used in theme three to allow a suitable geographical unit to enable at least ten deaths per MSOA for the survival calculations. Spatial patterns are observed in the data to determine the extent of clustering. Recommendations were made to the WCISU on completion of this research regarding the use of such clustering algorithms when analysing data at small area level, in particular for future cancer cluster enquiries that cancer registries receive and it is hoped that other cancer registries in the UK (and internationally) will adopt these recommendations. Appendix A shows the situation regarding the use of clustering algorithms by all cancer registries in the UK as of 2003. It can be seen that each cancer registry tends to adopt its own approach when dealing with such cluster enquiries. Hence, this research will hopefully provide some homogeneity between cancer registries in the UK when undertaking future cancer cluster enquiries.

Throughout the three themes, all clustering algorithms and analysis around point sources assumed a cluster that was circular in shape (for simplicity). In reality, this may not be the case. The postcode method is able to analyse an exposed area of any shape and a recent study by Kulldorff et al. (2006) (spatial scan statistic) has created an elliptical scan statistic.

Various existing clustering algorithms were examined for the first theme. There were a limited number of clustering algorithms available in ClusterSeer V2.2.4 and other software packages and algorithms could have been used. For example, DCluster is a suite of algorithms that can be used in the software R. However, ClusterSeer V2.2.4 was examined due to the ease of implementing the algorithms and the wide user base of these

tools. Confounding was a major issue in the first research theme since few algorithms in ClusterSeer V2.2.4 were able to take this into account. However, for themes two and three, confounding (age, sex and deprivation) was able to be taken into account due to the most effective algorithm from theme one being able to adjust for this major factor. Eight simulated datasets were used to determine the extent to which the clustering algorithms detected the artificial clusters. Many of these artificial clusters contained very high rates, much higher than the background rate; hence, the algorithms should have identified the clusters. Many more simulated datasets with varying degrees of clustering should be analysed using various clustering algorithms. The simulated datasets in past studies also appear to contain very high relative risks in the clusters which the clustering algorithms should have no trouble in identifying (Song et al., 2003). Various assumptions were made throughout the three themes which should also be taken into consideration, in particular when using the postcode method in theme two (3.6.3). The introduction of new digital products in recent years could have overcome some of these issues. Using a more recent dataset for the years of diagnosis, from 2000 onwards could also have overcome some of the issues raised in the research. For theme three, caution was advised when the number of deaths in a calculation was below 10, a rule that WCISU uses, due to the resulting survival estimates being unreliable. This problem was found for female breast cancer when analysing the separate two ten-year periods. If a later dataset was used using current data, for example, 1997-2006 then this problem should be reduced further due to the increasing incidence of breast cancer.

Previous studies have involved the use of clustering algorithms in an assessment of those that are more effective in terms of detecting actual clusters and when varying parameter levels but have tended to analyse just one or two clustering algorithms. The first theme has enabled a large number of algorithms to be compared using both real datasets (using leukaemia cases and population at risk data as well as leukaemia cases and control data) and simulated datasets. However, due to the type of clustering algorithm and data used by the algorithm, few comparisons could be made. Such studies generally tend to analyse case and control data since it is very difficult to obtain accurate population at risk data. This study has been able to analyse both types of data and the case and population at risk

data were further utilised in theme two with assumptions. It was concluded at WCISU that the current method of analysing data around hazardous sources using population weighted centroids had its limitations. The main reason for this was due to specific geographical units included within the exposed area were not being analysed due to the technique used. An improved method of identifying the specific area of interest was required, hence the methodological comparisons for the three techniques presented in theme two. It was concluded that the use of recent digital products for one of the techniques examined by other research groups was the more appropriate of the three techniques investigated since it only included those cases and population at risk within the exposed area of interest. Finally, the introduction of super output areas by the Office for National Statistics, in particular, MSOA has enabled relative survival analysis to be calculated at a small area level. This has not been possible to analyse in Wales in the past due to wards (908 using the 1991 census, 865 using the 2001 census) being too low a level to analyse survival and the next higher level geographical unit being Local Health Boards (22 in Wales) which were too coarse a geographical resolution to determine clusters of high and low survival in Wales. The introduction of these new levels of geography will permit a statistically reliable estimate of survival to be calculated and will also enable incidence and mortality of various cancers to be explored in further detail. These new levels of geography will also enable a temporal analysis if units are retained for the 2011 Census.

Future work

Following on from the results of this analysis, there are a number of potential avenues of research. These are described under each theme.

Theme one

A leukaemia dataset was used for the analysis in theme one because it was one of a number of cancers that are generally investigated when analysing small area level data

around hazardous sources (Waller et al., 1992; Coleman et al., 1989). This analysis could be repeated on another cancer that does not have a tendency to cluster to compare the results obtained with the leukaemia dataset used in this research. Additionally, only ClusterSeer and SaTScan were examined in this research. Another software package which is now available is DCluster that is implemented in "R". This package contains a number of general and focused clustering algorithms. Some of these algorithms are also available in ClusterSeer V2.2.4 so it would be beneficial to use the same algorithms and compare results with those presented in this thesis. It would also be useful to use any other clustering packages that have become available. When using two of the local methods, both required the user to input the maximum population size contained in any possible cluster. However, one algorithm varied the population size to be contained in a cluster whereas the other used this as the cluster size. Comparisons were made for various population sizes between the two methods. Further analysis could have been conducted to analyse the clusters obtained by one method in the context of the other. i.e. take a cluster obtained by one and analyse it using the other method in terms of the population size in the cluster to determine whether the same result would have been obtained.

More in-depth analysis of neighbouring clusters should also be undertaken due to the problems that the methods had when two clusters overlapped each other. The methods tended to produce an overall large cluster. Another of the algorithms located low and high value outliers in the leukaemia dataset. These should be examined further to determine if there was any reason as to why this was the case; for example was there a much higher population identified in this geographical unit compared to neighbouring areas? Alternatively, trends could be related to a nursing home or hospice situated in the geographical unit where many elderly patients may have been diagnosed or the area was more rural compared with its surrounding urban areas. These factors can be extended to the three themes. One of the algorithms produced conflicting p-values when increasing the number of nearest neighbours to analyse by just one. This should be explored further as to possible reasons as to why this was the case.

Finally, further research is warranted on simulated datasets to determine the algorithm that correctly detects the artificial clusters that have been introduced since only eight simulated datasets were explored here. This should involve many thousands of datasets of various cluster sizes and be used on point and areal data. The simulated datasets contained very high rates for the artificial clusters that were generated. Hence, the algorithms should have identified the artificial clusters. Other simulated datasets should be obtained to further test the clustering algorithms, especially the spatial scan statistic that was concluded to be the most effective algorithm from the first theme. Additionally, it would be useful to obtain case and control simulated datasets to analyse the clustering algorithms that only use case and control data since this was not investigated during the first theme.

There is scope for developing new clustering algorithms, or even extensions to existing clustering algorithms. For example, adjusting the algorithms in ClusterSeer V2.2.4 to account for confounding would be a major advantage. The method that details the most likely cluster for the maximum population size entered could be adapted so that it contains the most likely cluster from all possible population sizes for each centroid.

Theme two

Various assumptions were made when using the population based techniques to determine whether an increased risk existed. The product CodePoint, a database of all postcodes in the UK with various attributes was used on data in the 1980s and 1990s even though it was not released until 2000. The analysis should be repeated on more recent data due to the increasing digital products that are now available. A new version of CodePoint is distributed every three months. Hence, it would be useful to repeat the analysis on a cancer dataset for the period 2000-2006, the most recent year of diagnosis that data are published to see the effect of this. Another useful software product available from Ordnance Survey is AddressPoint. This software product contains all addresses in Great Britain to a resolution of 0.1 metres. Thus, this product could be used to compare

the results using the method identified in theme two as the best method to use with those using this new "address method" to determine if results are consistent when moving further down into a lower level of geography. ONS have also published head count data which contain populations for each postcode in England and Wales by sex for 2001 but with no age breakdown – this could be used as another method and compared with the postcode method to determine the comparability between methods. With the use of the ONS headcount data, the most effective algorithm could be rerun using postcode data as population data which is now available. However, the age distribution will not be able to be taken into account as a confounder since this information is not available. Assumptions, such as the same age distribution as its corresponding ED that the postcode was contained in, could be made regarding the distribution by age to enable age to be taken into account.

The analysis presented in this research was based on circular clusters. In the real world, clusters can be of any shape and size. Hence, the datasets could have been analysed using a recent algorithm created by Kulldorff (2006) that produces elliptical clusters. Additionally, the shape of the buffers could be changed to take into account the direction of exposure and the analysis repeated to determine if an increased risk exists. The circular clusters were used for simplicity but using attribute data for the landfill sites and local topographic as well as meteorological data it should be possible to, for example, develop more detailed exposure models.

Electric power line data were analysed as a potential hazardous source. Other linear hazardous sources such as rivers and railways were not taken into account here and should be investigated due to the large variation in results when analysing electric power line data with respect to the population at risk and specific cases that were included in the analysis using the centroid method and postcode method to determine whether the results obtained using electric power lines were an anomaly.

The latency period was investigated for just two clusters that the most effective clustering algorithm located in Wales; one cluster using a leukaemia dataset and the other using a

brain cancer dataset. It would be useful to determine the extent to which results differ when taking the latency period into account for all the landfill sites studied in this area of research in Wales as opposed to just two small clusters in Wales. It is postulated by Boer et al. (1997) and Pastor et al. (2005) that the deprivation distribution of the exposed population at risk is more biased towards the deprived at closer proximity to hazardous sources but is similar for all categories of deprivation a few kilometres away from the hazardous source. Thus, further work regarding the distance used and the age, sex, deprivation distribution from the hazardous source in question should be examined. In the past, categories of deprivation had been defined by an equal number of geographical units in each category of deprivation. However, as of May 2007, it was modified so that categories of deprivation be defined to be equal population or as close as possible in each category of deprivation. This was agreed at a United Kingdom Association of Cancer Registries (UKACR) Analysis Group in 2006, a group of analysts and statisticians from all cancer registries in the UK that meet four times a year. Thus, geographical units should be ordered by their deprivation score and allocated a category of deprivation to include a population size of $^{p}/_{c}$ where p is the total population size and c is the number of categories of deprivation. This was not done in this research since the analysis for this research had already been completed when the new method was initiated. However, when comparisons between these two methods were made at WCISU to calculate age standardised rates by deprivation quintile results were very similar when using equal geographical units in the quintiles or equal population in the quintiles.

This area of research was methodological rather than epidemiological in terms of comparing existing methods used to determine if an increased risk existed. Further analysis is warranted regarding use of home and work measurements in terms of exposure to determine if an increased risk exists around a specified hazardous source. Other factors not taken into account should also be considered such as stage of disease at diagnosis and direction of exposure.

Theme three

The types of survival analysis included as part of this theme should be repeated on more recent datasets since more data items are now collected by WCISU since this research was conducted. Treatment data, in terms of treatment type (surgery, chemotherapy or radiotherapy) and type of operation has been collected from the mid 1990s and stage of disease has also been collected. Thus, it would be beneficial to repeat the analysis using a 1994-2003 dataset allowing five years of follow up to the end of 2008 to see if there has been any change in the spatial patterns regarding relative survival from those examined here. Using this new dataset, stage of disease, and the distance from place of residence to place of treatment can be analysed to determine whether this affected the survival rates in rural areas.

Relative survival depends on the diagnosis date and death date of patients in a particular time period. Hence, it would be useful to map incidence and mortality by MSOA to see if the same spatial patterns were seen for incidence and mortality as were seen for survival, in terms of clusters of high and low survival since the low or high survival rates could be due to a very high incidence in a particular area or a very low mortality rate in a particular area.

Travel time analysis could be explored using a road network data set such as the Integrated Transport Network (ITN) in the dataset MasterMap. This details all road networks in the UK so that time of travel from place of residence to hospital can be determined. Relative survival by hospital could also be examined to determine those hospitals that perform better for particular types of cancer.

Higher Super Output areas are still to be defined by the Office for National Statistics. It may also be useful to calculate relative survival by the higher super output areas when they are released to determine if a similar pattern exists for female breast cancer and colorectal cancer as seen in the smoothing models.

The spatial scan statistic was used for the twenty year period examined for female breast cancer and colorectal cancer. The spatial scan statistic should be used on the ten year periods also examined to identify whether these clusters are consistent with those found when using Bayesian smoothing.

WCISU work closely with Screening Services in Wales and regularly exchange breast cancer and cervical cancer data between each other. It would now be possible to include those patients that have been diagnosed with breast cancer via screening into the model. WCISU are continually trying to increase the number of sources of information from which to get detailed data and obtain detailed pathology information from all but one of the pathology laboratories in Wales. This data were not available for the diagnosis years that were studied in this research. Radiology and hormonal treatment are not currently collected by WCISU but should be a priority to enable detailed treatment analysis to be carried out although it is acknowledged that there may be resource implications if this information is required to be entered manually.

Future work (all themes)

If the geography levels for the 2011 Census are the same as used for the 2001 Census, then the analysis should be replicated using this latest information for all three themes. Latest small area population figures will also be available following the 2011 Census which should be used in this latest analysis. Each of the themes could have benefited from detailed population estimation models but are confined to modelling estimated population from existing sources only.

Recommendations to cancer registries

Following this research, the following recommendations could be made to the WCISU and other cancer registries in the UK regarding small area analysis:

When a cluster enquiry is received by a cancer registry regarding a possible increased risk of a particular cancer around a point source, the spatial scan statistic should be used to identify a cluster for the cancer site in question using a suitable maximum population size (depending on the cluster enquiry and the extent of the cluster) to determine if the area at risk identified with the cluster enquiry is identified as a cluster. If this method identifies the area in question as a cluster, then the focused method should be used with the centroid of this cluster to identify if there is a particular time period whereby the cluster is significant. If the area of the cluster enquiry is not located as a cluster using the local method then no further analysis should be conducted. Even though the spatial scan statistic is fairly easy to use compared with other algorithms, it is advised that training should be provided to registries in the use of this technique.

Results show some population based techniques were very poor in terms of identifying the exposed area since cases and population at risk outside the exposed area were included in the analysis and cases and population at risk inside the exposed area were excluded from the analysis. Thus, these techniques should not be used for future analysis at cancer registries to identify whether an increased risk exists around a hazardous source. If these methods are used then the weaknesses in the methodology should be noted and included with the results of the analysis. The recently developed method by SAHSU should be used to identify whether an increased risk exists around a hazardous source since the geographical unit used does not affect the results and only those cases within the exposed area are analysed. This has the effect of not diluting the increased risk if a true increased risk exists, unlike the other methods. Also, the linear source analysis showed the large difference in the number of observed cases included in the analysis when comparing the postcode method with the centroid method. Sex-age specific rates for Wales are calculated to find the expected numbers in the exposed area. However, it is recommended that when calculating the background rate to find the expected numbers in the exposed area, the cases and population at risk inside the exposed area should be excluded from this calculation.

Additionally, the latency period should be taken into account, as well as the operation dates of the hazardous source, if known, since this can have an effect on the results. The latency periods for particular cancers should be reviewed through previous literature.

When examining relative survival rates at a low level of geography, the most effective algorithm from the first theme can be used to detect high and low rate survival clusters, if any exist, in the dataset. Regarding survival, it is interesting to note areas of worst survival (low rate clusters) as opposed to areas of high survival (high rate clusters). These clusters can then be explored in more detail as to reasons why such a cluster exists. Alternatively, smoothing should be used if areas of high and low survival rates are to be detected at a lower level of geography as it is difficult to identify areas of high and low survival at a small level of geography without any smoothing. Factors such as age, sex and deprivation should always be taken into account in any analysis along with other factors such as staging information if suitable data are available. The spatial scan statistic was the most effective at detecting clusters, if any existed and should be used over the majority of other algorithms examined in this research. Another clustering algorithm was useful in identifying outliers which can then be explored in further detail. The introduction of MSOA have allowed relative survival estimates to be calculated at small area level in Wales whereas in the past, reliable survival estimates at small area level was not capable before this new geographical level was defined. This research is timely in that it uses new digital products that have recently come online. Additionally, further extensive ranges of these digital products are now available that were not available at the beginning of this research and can be used to produce new and innovative improvements to methods detailed in this research.

There are particular data items that would have been useful to collect in order to perform analysis when taking into account these additional factors such as stage of disease and treatment information. These items were poorly collected for the diagnosis years that were analysed here although recent data suggest that these data items are more populated than previous years. Joint working with other organisations such as the Environment Agency or the Health Protection Agency is beneficial and advisable to overcome

particular hurdles in obtaining particular datasets such as the hazardous sources, as without this information, the analysis would not have been possible.

In conclusion, it is anticipated that together these research studies will be of wider benefit to other registries in the UK charged with investigating spatial and temporal variations in cancer rates. In addition, the recommendations to cancer registries highlighted here can certainly be extended to those throughout Europe and even other cancer registries in the world if suitable geographical units with reliable population estimates are available.

2. THEME ONE

Comparison of cluster detection tests using cancer datasets in Wales

In epidemiological terms, a cluster is an aggregation of cases of a disease closely grouped in time, in space or in both time and space such that the number of cases found in the cluster is much greater than the number of cases that would have been expected. Alternatively, the number of cases found in a cluster could be much lower than the number of cases that would have been expected giving rise to a "low value" cluster as opposed to a "high value" cluster. A large number of clustering techniques are currently available to use in order to evaluate whether a cluster of a specific disease exists. The problem is that many of these algorithms will provide different results due to the technique used and the data requirements.

2.1. Aims and objectives

The main aims and objectives for the first theme were as follows:

- To examine the various clustering algorithms included in the software package ClusterSeer V2.2.4 and SaTScan V5.1.3 using real data sets and simulated datasets of case and population at risk data or case and control data.
- Compare results of the clustering algorithms (for those which could be compared to each other) to determine the most effective algorithm when dealing with cancer cluster enquiries at cancer registries in the UK.

The main aim of this area of research was to investigate the various clustering algorithms that were included in the software package ClusterSeer V2.2.4, produced by BioMedware which consists of a large number of clustering algorithms, and SaTScan V5.1.3[1] (Kulldorff, 1997, 2004), a software package that is freely available on the Internet and released on 18th April 2005. ClusterSeer V2.2.4 was examined due to the vast number of algorithms that the software houses. Real datasets are used to compare the algorithms. This theme examines spatial clustering and space-time clustering algorithms only. Algorithms for investigating temporal clustering are available but are not utilised here. It

[1] www.satscan.org/

was not the purpose to derive a new algorithm in this research but to provide a brief overview of each method, along with its strengths and weaknesses, to determine a test that should be used for the second and third stages of work. Recommendations were made regarding the wider use of each clustering algorithm along with an overall aim to make recommendations to WCISU regarding the use of algorithms for cancer cluster analysis.

This theme took different scenarios and compared results between specific clustering algorithms. A number of simulated datasets were also used to determine whether the algorithms correctly detected the artificial clusters that were introduced into the datasets. However, as stated earlier only a limited number of simulated datasets were examined here.

Input parameters were examined in greater detail to explore whether this affected the resulting p-values and conclusions and to identify if particular algorithms performed better than others.

A major advantage that this research has over previous work is that a large number of clustering algorithms were analysed using the same datasets. The results were compared and contrasted for those algorithms that were comparable. The datasets also had the advantage of better accuracy of postcode information. Research to date has used accuracy of postcode data to 100 metres. However, with the use of the Ordnance Survey product CodePoint^{TM}, postcodes have grid references (average of all homes with the same postcode) accurate to one metre.

To enable homogeneity between cancer registries, this research will deliver recommendations to all cancer registries in the UK when dealing with cluster investigations. This initial area of research led to the second and third areas of research.

2.2. Introduction to clustering

The study of disease clusters may show possible factors and exposures that influence risk for a particular disease. Substantial advances in computer technology such as Geographical Information Systems (GIS) in recent years have enabled the integration of statistical techniques which offer new and more accurate methods of analysing disease clusters. There are three general types of clustering:

- Spatial Clustering
- Temporal Clustering
- Space-Time Clustering

Besag and Newell (1991) classified spatial clustering into general spatial clustering and focused spatial clustering. Focused methods detect clusters around a particular point source(s). General methods can be classified further into global and local methods of clustering. Global methods examine the dataset to identify any unusual spatial patterns. Global methods do not give locations of any clusters in the dataset; they summarise whether the entire region shows evidence of clustering or not. Local methods are used to identify the location of clusters in the study region. i.e. lists of clusters, if any, are identified in specific locations. Temporal clustering detects clusters in a particular time period. Space-time clustering identifies clusters in particular areas of the study region that depend on the time period. Each clustering algorithm can be described using the following structure.

- **Null and Alternative Hypothesis**

  The null hypothesis states that there is no clustering in the dataset being examined. The alternative hypothesis is defined as clustering and may include parameters to define spatial, temporal or space-time clustering.

- **Test Statistic**

  The test statistic summarises the data with reference to the hypotheses that are of interest. The null distribution of the test statistic is calculated empirically by

Monte Carlo Randomisation or from distributions such as the Poisson distribution. Monte Carlo Randomisation is a method whereby the observations from the original dataset are randomised using various techniques described later and the test statistic is recalculated for each randomisation. It essentially computes the distribution under the null hypothesis Ho. There are various randomisation techniques that can be used depending on the type of clustering algorithm. The randomisations are repeated many times amassing distributions to calculate a p-value. P-values are calculated by comparing the observed test statistic to the null distribution.

- **P-values**

  The test statistic among the Monte Carlo randomisations is ranked to obtain a p-value. The formula for an upper tailed p-value is $P_{upper} = \frac{MCR_1 + 1}{MCR + 1}$ where MCR is the total number of Monte Carlo randomisations and MCR_1 is the number of Monte Carlo Randomisations in which the statistic was more extreme than or equal to the observed statistic.

- **Monte Carlo Randomisation techniques**

  There are various randomisation techniques used by clustering algorithms in ClusterSeer V2.2.4 to determine the p-value. The type of randomisation technique used depends on the specific clustering algorithm. These methods include: generating randomised case counts from a *Multinomial distribution or Poisson distribution, conditional randomisation* (the disease frequency in one location is fixed and the others are randomly assigned new locations for each randomisation), *randomised data* (each location is assigned data from another location), changing distances between points by multiplying by a *random number*, *shuffling time distances* while keeping the spatial distances constant or *shuffling the time occurrence* of each case between other cases and keeping the locations of the cases fixed.

- **Adjustments for multiple comparisons**

  It is sometimes necessary to adjust the level of significance since a statistical test may be performed multiple times on the same dataset (and hence an increase in the likelihood of wrongly rejecting the null hypothesis when it is true can occur). To allow for this, ClusterSeer V2.2.4 lowers the significance level using various methods. These methods are shown in equation 2.1. α is the original significance level, j is the number of comparisons carried out at the initial significance level. The Simes and Holm's adjustments are performed for each test, sequentially ordered from lowest to highest p-value, with i denoting the sequential index (range: 1..j) for the individual test considered. Thus, the adjusted significance levels in an individual comparison will be less than or equal to the unadjusted significance level, α.

$$Bonferroni = \frac{\alpha}{j} \qquad \text{Bland et al (1995)}$$

$$Sidak = 1-(1-\alpha)^{\frac{1}{j}} \qquad \text{Sidak (1967)}$$

$$Simes = \frac{i\alpha}{j} \qquad \text{Simes (1986)}$$

$$Holms = 1-(1-\alpha)^{\frac{1}{j-1+i}} \qquad \text{Holland et al (1987)}$$

Equation 2.1: Adjustments of significance level in ClusterSeer.

ClusterSeer V2.2.4 additionally calculates combined p-values for all tests performed at the initial significance level for Bonferroni and Simes adjustments and these are shown in table 2.1 where P_i is the p-value of an individual test.

Adjustment	Combined p-value
Bonferroni	$j*\min(P_i)$
Simes	$\min[(j-i+1)*P_i]$

Table 2.1: Combined p-values for Bonferroni and Simes adjustment.

Software

ClusterSeer V2.2.4 is a software package available from BioMedware that comprises twenty five clustering algorithms that can be implemented to analyse the level of clustering. This package was chosen since it contained many of the algorithms used in previous studies. There are various types of clustering that the algorithms are able to analyse. Some cluster algorithms require case and control data. This can be aggregated to a specific geographical unit or individual locations. Other algorithms require case and population at risk data while others use case data only which again can be aggregated to a specific geographical unit or individual locations. However, population data must be available at the specific geographical unit used to analyse such data. Table 2.2 details the type of detection associated with the clustering methods in ClusterSeer V2.2.4. The freely downloadable software package SaTScan V5.1.3 (Kulldorff, (1997, 2004)) has also been used to identify local and focused clusters in previous studies.

Spatial Cluster Detection	Temporal Cluster Detection	Space-Time Cluster Detection	Surveillance
Anselin's Local Moran	Dat	Direction	Rogerson
Besag and Newell	Ederer-Myers-Mantel	Grimson	
Bithell's Linear Risk Score	Empty Cells	Jacques k-NN Nearest Neighbour	
Cuzick and Edwards	Grimson	Knox	
Diggle	Larsen	Kulldorff	
Getis-Ord Local G	Levin and Kline	Mantel	
Grimson	Wallenstein		
Kulldorff			
Moran's I			
Oden's I Pop			
Ripley's K function			
Score Test			
Turnbull			

Table 2.2: Clustering Algorithms available in ClusterSeer.

Table 2.3 illustrates a typology of clustering algorithms and the data format that is required to run the algorithms in ClusterSeer V2.2.4 and SaTScan V5.1.3 and table 2.4 gives a summary of the type of data that is required, whether the algorithm can adjust for confounding in ClusterSeer V2.2.4 (this occurs when the apparent effect of exposure on disease is distorted by a confounding factor which is associated with both the disease and

the exposure), the Monte Carlo randomisation technique and examples of studies where the particular algorithm has been used in previous research.

Clustering Algorithms in ClusterSeer and SaTScan	Global Methods	Local Methods	Focused Methods	Space-Time Methods
Case and Population At Risk	Besag and Newell Oden's I Pop	Besag and Newell Kulldorff[1, 2] Turnbull	Bithell's Linear Risk Score Score Test Kulldorff[1]	Kulldorff[1, 2]
Case and Control	Cuzick and Edwards	Kulldorff[1]	Diggle Kulldorff[1]	Kulldorff[1]
Rates	Moran's I			
Cases Only	Grimson Ripley	Getis-Ord Local Moran		Direction Grimson Jacques Knox Mantel

1 Available in SaTScan 2 Available in ClusterSeer

Table 2.3: Data requirements for the different types of clustering algorithms.

Many of the clustering algorithms detailed in table 2.3 and table 2.4 use case data only but Wales is a country of varying population size between neighbouring geographical units. i.e. the population distribution is not constant throughout Wales. Thus, these algorithms will not perform as they should and results will be invalid. These algorithms will not be explored in any more detail; those being Grimson's method, Ripley's method, Getis-Ord method, Direction method, Jacquez's method, Knox method and Mantel's method. Anselin's Local Moran test is not a clustering algorithm per se but a test to determine if neighbouring areas are similar or dissimilar and can identify outliers. Thus this method will be included in the following analysis to determine if outliers are present in the data. If there are any identified outliers, these can be explored in further detail.

Cluster Algorithm	Data	Confounding	MC randomisation technique	Examples
Moran's I Statistic	Grouped level - rates	No	Randomised data	Cullen et al (2001), Castresana et al (2002)
Oden's I Pop	Grouped level - case and population	No	Multinomial	Fosgate et al (2002)
Besag and Newell	Grouped level - case and population	No	Multinomial	Besag and Newell (1991), Waller et al (1994)
Cuzick and Edwards	Individual level - case and control	Yes	Randomised data	Cuzick and Edwards (1990), Dockerty et al (1999)
Ripley's K Function	Individual level - case	No	Random number	Grau (2002)
Kulldorff's Spatial Scan Statistic	Grouped level - case and population (or case and control)	Yes*	Multinomial	Hjalmars (1996), Kulldorff et al (1997)
Anselin's Local Moran Test	Grouped level - case	No	Conditional	Jacquez and Grieling (2002)
Turnbull's Method	Grouped level - case and population	No	Multinomial	Turnbull et al (1990)
Getis-Ord G Statistic	Grouped level - case	No	Randomised data	Jeffery et al (2002), Ceccato and Persson (2002)
Score Test (Lawson and Waller)	Grouped level - case and population	No	Poisson	Waller et al (1992)
Bithell's Linear Risk Score Test	Grouped level - case and population	No	Poisson (Unconditional)	Bithell (1995)
Diggle's Method	Individual level - case and control	Yes	Maximum likelihood estimation	Diggle (1990), Diggle and Rowlingson (1994)
Jacquez's k-NN Method	Individual level - case	No	Shuffling time distances	Norstrom et al (2000), Van Buuren et al (1998)
Knox Method	Individual level - case	No	Shuffling time distances	Gilman et al (1999), Machado-Coelho et al (1999)
Mantel's Method	Individual level - case	No	Shuffling time distances	Schmucki et al (2002)
Direction Method	Individual level - case	No	Shuffling time occurrence	Jacquez et al (1994)

** Only available in SaTScan*

Table 2.4: Summary of clustering algorithms

Geographical software used in the current research includes ArcView V3.2a and ArcGIS V8.3. The extension ProAddress is used in ArcGIS V8.3 to update the postcode of residence at diagnosis to the corresponding current postcode. CodePoint^{TM} (software available from Ordnance Survey) allocates the updated postcode to a northing and easting to a resolution of one metre. The northing and easting for a postcode is the average of all eastings and northings for all addresses of a particular postcode.

Table 2.5 shows a random sample of 447 coordinates generated in 16 neighbouring wards in Wales using resolutions of one metre and one hundred metres to determine which ward the coordinate was placed. As can be seen, many coordinates moved from one ward to

another ward and four cases moved to completely different wards when the accuracy of the geocoding decreased. In total 32 (7.2%) of the coordinates changed wards. Thus this study uses the most accurate geocoding possible. Since this study uses small area data, an increase of just one or two cases can dramatically affect results.

Ward	1m	100m
Ward 1	12	13
Ward 2	29	28
Ward 3	31	28
Ward 4	34	35
Ward 5	29	29
Ward 6	49	42
Ward 7	19	22
Ward 8	37	35
Ward 9	14	12
Ward 10	19	21
Ward 11	23	25
Ward 12	27	27
Ward 13	41	42
Ward 14	28	29
Ward 15	22	21
Ward 16	33	34
Other wards	-	4

Table 2.5: Analysis of 1m and 100m coordinates.

2.3. Description of case datasets and control datasets

WCISU is an organisation located in Cardiff, South Wales that receives registrations for all cancers diagnosed in residents of Wales from 1974 to the present day. Cancers are coded using the 10th edition of the International Classification of Diseases (ICD10) from 1995 to the present day. ICD9 codes were used for the diagnosis years 1974-1995. Cancer registrations are received via a number of sources, primarily from PEDW (Patient Episode Database for Wales). Other sources include pathology, death certificates and specialist cancer databases. Following ethical approval the cancer data acquired from the WCISU was aggregated into a non-identifiable format for use in this research.

It is known that leukaemia rates for males and females in Wales are higher than in other parts of the United Kingdom and Ireland. It is thought that there is an inverse association

between incidence and socio-economic deprivation but none of the known risk factors could explain recent observed geographical variations in the latest report from the Office of National Statistics regarding leukaemia in the UK and Ireland for the period 1991-1999[2].

Previous studies tend to look at cancers such as leukaemia, childhood cancer and brain cancer when determining whether clusters are located in particular areas in relation to hazardous sources (Waller et al. (1992), Coleman et al. (1989), Feychting et al (1994)). Therefore, for this theme, leukaemia was investigated due to the larger number of cases in Wales compared with the other two types of cancer.

A dataset on all leukaemias in Wales for the period 1982-2001 was extracted from the WCISU database to evaluate the clustering methods. Wards defined from the 1991 UK Census were used since these corresponded to the centre of the diagnosis period studied. Ward population figures were obtained from Census Area Statistics on the Web (CASWEB) using the 1991 UK Census. Boundary files of wards in Wales defined by the 1991 UK Census were obtained from the website UKBorders[3].

A dataset of leukaemia cases (ICD 9 codes 204.0-208.9, ICD10 codes C91.0-C95.9) was extracted from the WCISU database to evaluate the clustering methods. The ICD9 codes and ICD10 code definitions are shown in table 2.6.

ICD9	ICD10	Definition
204.0-204.9	C91.0-C91.9	Lymphoid leukaemia
205.0-205.9	C92.0-C92.9	Myeloid leukaemia
206.0-206.9	C93.0-C93.9	Monocytic leukaemia
207.0-207.9	C94.0-C94.9	Other leukaemias of specified cell type
208.0-208.9	C95.0-C95.9	Leukaemia of unspecified cell type

Table 2.6: ICD9 and ICD10 code definitions for leukaemia.

[2] http://www.statistics.gov.uk/downloads/theme_health/caUKI91_00/Ch11_Leukaemia.pdf
[3] http://edina.ac.uk/ukborders/

There were 7689 cases of leukaemia in Wales for the period 1982-2001; 4316 cases of those were males (3.1% of all male malignancies excluding non-melanoma skin cancer) and 3373 females (2.4% of all female malignancies excluding non-melanoma skin cancer). Crude rates for male leukaemia have increased from 10.1 per 100,000 population in 1982 to 17.6 per 100,000 population in 2001; this is shown in figure 2.1.

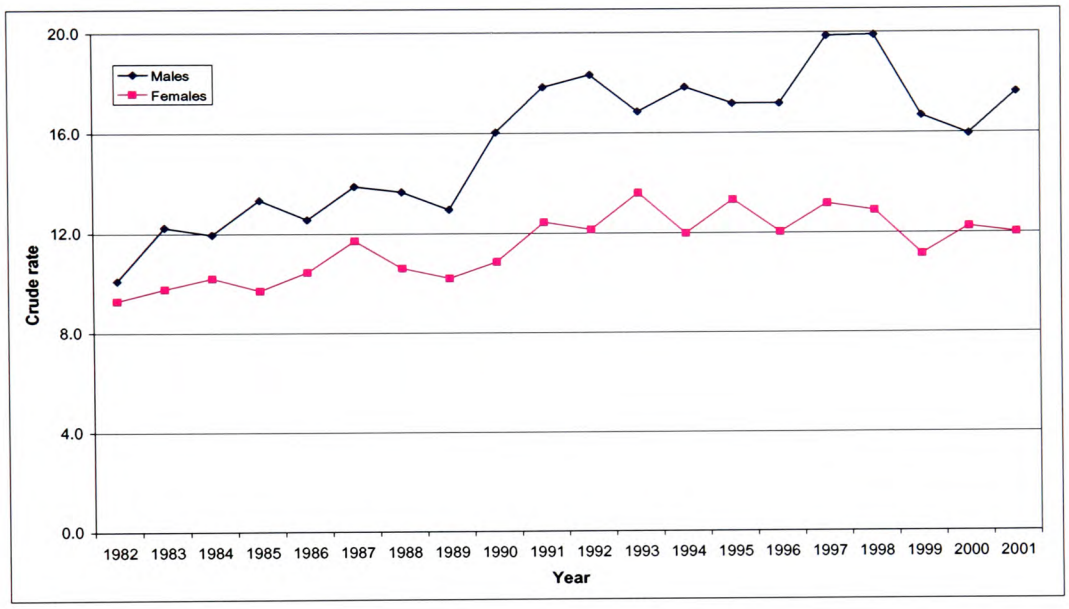


Figure 2.1: Crude rates per 100,000 population for leukaemia in Wales 1982-2001.

Female leukaemia has increased from a crude rate of 9.3 per 100,000 population in 1982 to 12.0 per 100,000 population in 2001, an increase of nearly 30%. Childhood leukaemia (0-14 years) accounted for over 5% of all leukaemia cases for males and approaching 6% for all female leukaemia cases. The mean age of diagnosis ranged from 61.0 years (in 1990) to 65.8 years (in 1997) for males and from 61.6 years (in 1999) to 69.8 years (in 1997) for females for the twenty year period 1982-2001.

In the current study of the 7689 cases of leukaemia diagnosed in the period 1982-2001, 76 cases could not be allocated an enumeration district using the 1991 census data due to a missing postcode at diagnosis. The majority of these cases were diagnosed in the initial ten-year period 1982-1991. Cancer tends to be associated with higher levels of deprivation, thus this factor should be accounted for in the analysis. The Townsend score

(Townsend et al., 1988) is a measure of area deprivation. It is made up of four variables: unemployment, car ownership, owner occupation and overcrowding. All enumeration districts in Wales were assigned a quintile of deprivation based on the Townsend score such that each quintile contained approximately the same number of enumeration districts, thus the data can then be adjusted for deprivation from affluent (quintile 1) to deprived (quintile 5). Each case was assigned their enumeration district of residence and resulting Townsend quintile. Of the remaining cases in the analysis, 2 cases could not be allocated a Townsend quintile based on enumeration districts using the 1991 census data. In total, 78 cases were not included in the analysis, giving a revised total of 7611 cases of leukaemia investigated. Each leukaemia case was matched with two controls by five year age band, sex and deprivation quintile from an extract of the National Health Service Administrative Register (NHSAR) in 1997; a listing of all patients registered with a general practitioner (GP). Two controls per case were used to enable a large number of controls to reflect the population of Wales. Two controls per case were selected from the NHSAR for this analysis; however, three, four or even five controls per case could have been selected to reflect the background population at risk. The age at diagnosis of each case was revised to their age in 1997 when obtaining the controls. For example, a woman aged 54 (age band 50-54), diagnosed with leukaemia in the year 1985 and living in an area where the Townsend quintile of deprivation was classed as 2 would have been matched with a woman aged 66 (age band 65-69) and Townsend quintile 2 so that the control would have been the same age at diagnosis as the case. Table 2.7 shows the distribution of cases by sex and Townsend quintile of deprivation along with the crude rates and age-deprivation standardised rates (per 100,000 population) for males and females.

MALES					
	TQ1	TQ2	TQ3	TQ4	TQ5
Cases	837	831	830	919	864
Crude Rate*	13.97	14.89	15.74	18.16	16.13
ADSR*	15.60	14.26	14.96	17.14	17.05
FEMALES					
	TQ1	TQ2	TQ3	TQ4	TQ5
Cases	599	630	676	724	701
Crude Rate*	9.18	10.39	11.94	13.46	12.50
ADSR*	10.31	9.68	11.18	12.91	13.98

ADSR: age-deprivation standardised rate * per 100,000 population

TQ1: affluent, TQ5: deprived

Table 2.7: Age standardised rates for leukaemia in Wales 1982-2001.

As can be seen from table 2.7, there appears to be no apparent trend in the crude rates (CR) per 100,000 population by Townsend quintile of deprivation (TQ) in Wales for the period 1982-2001 for males or females (TQ1 represents affluent, TQ5 represents deprived). The age-standardised rates show a generally increasing trend in leukaemia rates towards the deprived group (if excluding the affluent group, TQ1).

Lawson (1999) states the risk of leukaemia can be sensitive to higher levels of airborne environmental pollution. Lower body cancers such as prostate, testes, cervix and uterus can be considered as controls due to their lack of known correlation with air pollution (Lawson, 2001, p156). The lower body cancers also have a similar age structure to the leukaemia cases. Thus a dataset of lower body cancers in Wales for the same time period was used as an alternative control dataset for the leukaemia cases for the case-control cluster methods. There were 31,895 cases of lower body cancer in Wales for the period 1982-2001; 33 cases were unable to be matched with a Townsend quintile of deprivation and 80 cases had a missing postcode and were not included in the analysis giving a total of 31,782 controls in the analysis. Table 2.8 shows the distribution of cases for leukaemia and lower body cancers in Wales for the period 1982-2001.

Age Band	0-4	5-9	10-14	15-19	20-24	25-29	30-34	35-39	40-44	45-49
Leukaemia	3.0	1.4	1.2	1.4	1.2	1.2	1.6	2.1	2.2	3.0
LBC	0.0	0.0	0.0	0.2	0.6	1.4	2.2	2.6	2.3	2.3

Age Band	50-54	55-59	60-64	65-69	70-74	75-79	80-84	85+	Total
Leukaemia	3.4	5.7	9.0	12.0	14.0	14.2	12.5	10.7	100.0
LBC	3.5	5.8	9.2	13.6	17.1	17.2	12.7	9.3	100.0

Table 2.8: Proportion of cases by five year age band for leukaemia and lower body cancers in Wales, 1992-2001.

Thus for case and control methods, the leukaemia cases and the two sets of controls (NHSAR and lower body cancers) were used to compare algorithms as well as comparing both sets of control results to identify if they both produce similar results. For case and population at risk methods, the leukaemia cases and the corresponding population at risk were used in the analysis. Note that the format of the data files for each algorithm differs slightly between algorithms. Care must be ensured when preparing the data for each algorithm due to the range of formats used.

2.4. Global Clustering Methods

2.4.1. Moran's I Statistic

Moran's I statistic (Moran 1950) is a global measure that detects departures from spatial randomness and requires rate data for each aggregated geographical unit e.g. wards, enumeration districts.

Test statistic

Spatial patterns are identified via departures from spatial randomness. Neighbouring areas that have similar rates indicate positive spatial autocorrelation and global spatial clustering. Moran's I test statistic is shown in equation 2.2. n represents the total number of areas in the geography base i.e. the total number of wards in Wales using the 1991 census data. For the following analysis this value was 908. w_{ij} represents the weight that

denotes the strength of the connection between areas i and j. z_i represents the difference between the rate in area i and the average rate for the entire study region.

$$I = \frac{n\sum_{i=1}^{n}\sum_{j=1}^{n,j\neq i} w_{ij} z_i z_j}{\sum_{i=1}^{n}\sum_{j=1}^{n,j\neq i} w_{ij} \sum_{i=1}^{n} z_i^{\,2}}$$

Equation 2.2: Moran's I test statistic.

The test statistic is positive when neighbouring wards have similar rates and negative when neighbouring wards have dissimilar rates. The expected Moran statistic is given by $E(I) = -\frac{1}{(n-1)}$ which is -0.01103 for all five-year periods. Thus, for smaller geographical units, the closer to zero the statistic is expected to be. This method reports a two tailed p-value as this statistic detects whether Moran's I is positive or negative. The variance is determined under the normal assumption or randomisation assumption. The variance is shown in equation 2.3 under the randomisation assumption along with the calculation of the z-score to determine significance. This is the method that is usually used when dealing with disease rates.

$$Var(I) = \frac{n[(n^2-3n+3)a - nb + 3f^2] - \frac{g}{h^2}[(n^2-n)a - 2nb + 6f^2]}{(n-1)(n-2)(n-3)f^2} - E(I)^2$$

$$z = \frac{I - E(I)}{\sqrt{Var(I)}}$$

$$\text{where } a = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n,j\neq i}(w_{ij} + w_{ji})^2,\ b = \sum_{i=1}^{n}(w_{i*} + w_{*i})^2,\ f = \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}, j \neq i$$

$$g = \frac{1}{n\sum_{i=1}^{n} z_i^{\,4}},\ h = \frac{1}{n\sum_{i=1}^{n} z_i^{\,2}}$$

Equation 2.3: Variance and calculation of z-score for the Moran Test under the randomisation assumption.

The significance of Moran's I is evaluated by Monte Carlo simulations and via the z-score above.

Two options are available for the adjacency of polygons when using this algorithm. The "rook" method identifies polygons that surround a particular polygon of length greater than zero (i.e. more than one point in common) whereas the alternative "queen" method identifies polygons that surround other polygons even if the polygons have just one single point in common. Figure 2.2 displays the difference between these two methods. Both diagrams show a number of polygons. The black polygon is the polygon in question. Under the rook method, all polygons highlighted in grey are classed as neighbours of the black polygon. The same has been done for the queen method. The queen method has one extra neighbour, polygon "A". It is adjacent to the black polygon by a single point, thus the rook method will not incorporate this polygon into the analysis.

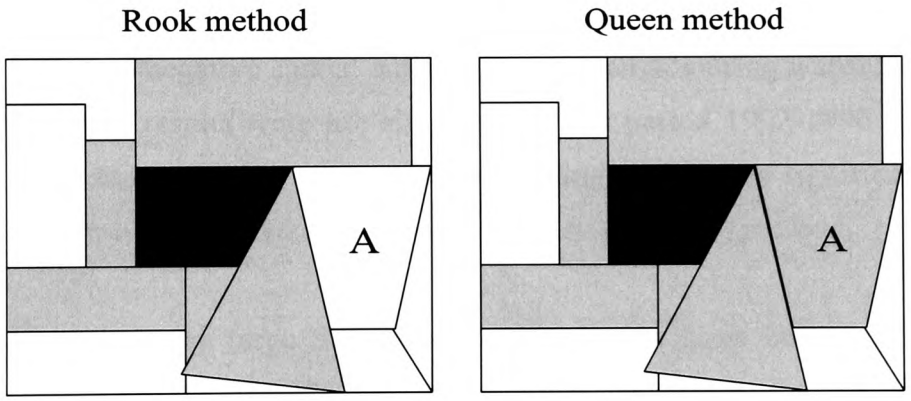


Figure 2.2: Rook contiguity and queen contiguity.

Results

Ho: Moran's I is zero. Disease rates are spatially independent.

Ha: Moran's I is not zero. Disease rates are not spatially independent.

Using the dataset of leukaemia cases in Wales for the period 1982-2001 the twenty year period was split into four five-year periods. The results using the rook and queen contiguity methods are detailed in table 2.9.

	1982-1986	1987-1991	1992-1996	1997-2001
Average disease frequency*	10.405	12.775	15.167	15.343
ROOK CONTIGUITY	**1982-1986**	**1987-1991**	**1992-1996**	**1997-2001**
Moran's I value	-0.007	-0.011	0.046	-0.015
Monte Carlo p-value	0.744	0.674	0.020	0.488
QUEEN CONTIGUITY	**1982-1986**	**1987-1991**	**1992-1996**	**1997-2001**
Moran's I value	-0.005	-0.012	0.048	-0.014
Monte Carlo p-value	0.908	0.606	0.016	0.588

* *per 100,000 population*

Table 2.9: Analysis of leukaemia dataset using Moran's I method.

As can be seen from table 2.9, the output lists the overall background rate per 100,000 population. This figure steadily increases from 10.405 per 100,000 population at the earliest period to 16.343 per 100,000 population at the latest period. Moran's I test statistic was negative (and very similar to that expected) for all periods excluding 1992-1996. This indicated negative spatial autocorrelation; neighbouring wards had dissimilar rates, although these results were not significant. The period 1992-1996 indicated that neighbouring wards had similar rates, this result being statistically significant at the 5% level for both the rook ($p=0.020$) and queen contiguity ($p=0.016$) method.

This method is biased by large differences in population sizes between neighbouring areas since the population of each ward is not taken into account. The population by ward in Wales differs largely between neighbouring areas, hence Oden's I Pop method (2.4.2) should be used if population at risk data are available.

2.4.2. Oden's I Pop Method

Oden (1995) adjusted Moran's I statistic to account for population differences across areas. This method required case and population data at group level i.e. wards. Areas

with large differences in population size decrease the power of Moran's I statistic to identify a true cluster or departure from spatial randomness.

Test statistic

Oden's I Pop Statistic (*Ipop*) is shown in equation 2.4 where c_1 is the total number of cases in the study region, n is the total number of wards, c_i is the total number of cases in ward i, d_i is the proportion of the population in ward i, p^* is the total population at risk in the study region, w_{ij} is a weight denoting the connection between ward i and ward j, that is 1 if they share a common border and 0 otherwise.

$$Ipop = \frac{c_1^2 \sum_{i=1}^{n}\sum_{i=1}^{n} w_{ij}\left(\frac{c_i}{c_1} - d_i\right)\left(\frac{c_j}{c_1} - d_j\right) - c_1\left(1 - \frac{2c_1}{p^*}\right)\sum_{i=1}^{n} w_{ij}\frac{c_i}{c_1} - \frac{c_1^2}{p^*}\sum_{i=1}^{n} w_{ii}d_i}{S_0 \frac{c_1}{p^*}\left(1 - \frac{c_1}{p^*}\right)}$$

$$S_0 = p^{*2}\, A - p^* B,\ A = \sum_{i=1}^{n}\sum_{j=1}^{n} d_i d_j w_{ij},\ B = \sum_{i=1}^{n} d_i w_{ii}$$

Equation 2.4: Oden's I pop statistic.

The expected value for Oden's I statistic is $E(Ipop) = -\frac{1}{(p^*-1)}$ where p^* is the total population at risk for the entire study region. Hence, the expected value approaches zero for large populations. The output also produces an adjusted statistic for *Ipop* based on the average rate throughout the entire study region and is calculated by $Ipop' = \frac{(p^*)Ipop}{n}$ where n is the total number of wards in the study region. As with the previous method the option of rook or queen contiguity is given. In this study, both have been used to compare results.

Results

Ho: Oden's I pop is zero. Disease rates are spatially independent.
Ha: Oden's I pop is not zero. Disease rates are not spatially independent.

The leukaemia dataset for the period 1982-2001 was analysed using this method for the same four five-year periods as previously studied. Table 2.10 shows a summary of the results using Oden's method.

	ROOK METHOD				QUEEN METHOD			
	1982-1986	**1987-1991**	**1992-1996**	**1997-2001**	**1982-1986**	**1987-1991**	**1992-1996**	**1997-2001**
Ipop	$9.3*10^{-6}$	$9.6*10^{-7}$	$1.1*10^{-5}$	$6.1*10^{-6}$	$9.6*10^{-6}$	$9.6*10^{-7}$	$1.1*10^{-5}$	$6.1*10^{-6}$
Ipop'	0.029	0.003	0.034	0.019	0.030	0.003	0.034	0.019
% within	98.333	102.395	96.024	100.531	97.500	102.590	95.545	100.187
% among	1.667	-2.395	3.976	-0.531	2.500	-2.590	4.455	-0.187
p-value	0.014	0.357	0.001	0.018	0.014	0.344	0.001	0.017

Table 2.10: Analysis using Oden's method.

For all periods, Oden's I pop method produced a statistic very close to zero. The values of the standardised statistic (Ipop') also show values close to zero indicating consistency with spatial randomness. Excluding the period 1987-1991, all results are significant at the 5% level. The proportion of the test statistic that is attributable to clustering with the study region (Wales) is denoted by "% within" and the proportion of the test statistic that accounts for the excess of cases being similar in neighbouring wards is denoted by "% among". If this value is negative, it implies that there is dispersion of cases in adjacent areas. Figure 2.3 shows the histograms obtained for each period along with the value of the test statistic (thick black line) for the rook method. The histograms show the reference distribution generated by randomising the dataset and recalculating the statistic for the randomised data. The histograms support the evidence obtained in table 2.10.

Comparing table 2.10 with that of Moran's I results in table 2.9, the periods 1982-1986 and 1997-2001 are also significant using Oden's I Pop method. Only the period 1987-1991 remains non-significant. The p-value for the period 1992-1996 is highly significant

at 0.001 (compared with 0.016 using Moran's I method). These differences are due to Oden's method using the population of Wales for each ward whereas Moran's method does not take population count into account – it is taken into account in the rates but does not determine whether the number of cases and population data per ward is large or small. Thus, Oden's method should be used instead of Moran's method if population at risk is available.

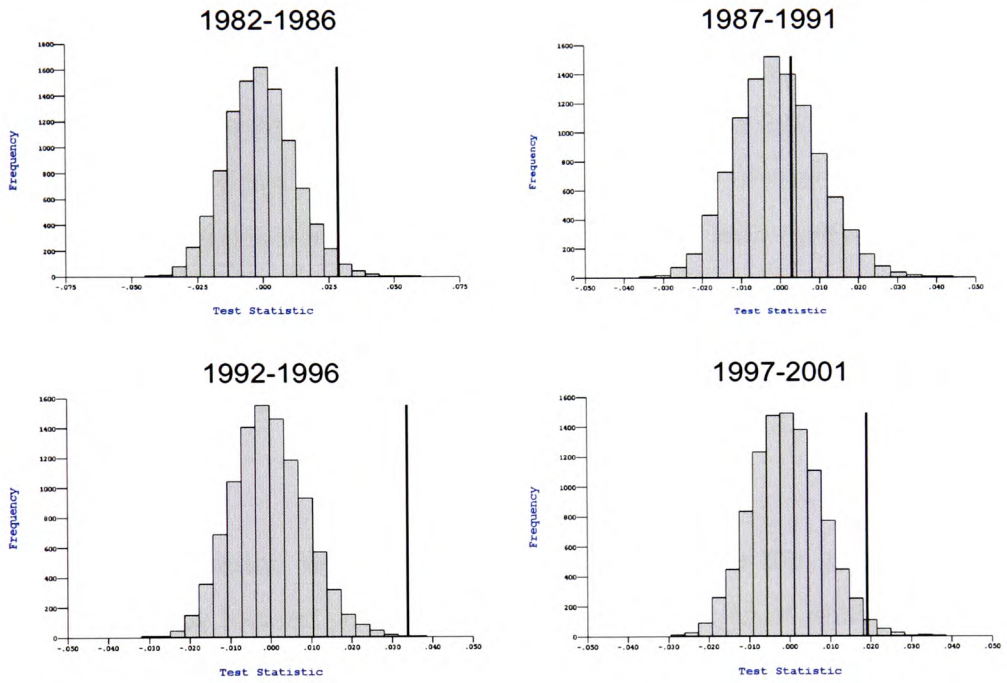


Figure 2.3: Histograms using the leukaemia dataset for Oden's I pop method, rook method.

2.4.3. Besag and Newell's Method

Besag and Newell (1991) proposed an algorithm that used case data and population at risk data for each aggregated area. A circular window is centred on each area in turn and expanded to include adjacent areas until the total number of cases in the circular window reaches a predetermined maximum cluster size that is input by the user. The population size inside the circular window is compared with that expected for the whole of Wales. This test is ideally suited for small population totals that are aggregated by area (ward). This method is both a global and local cluster detecting algorithm. Global clustering is presented here; local clustering is presented in section 2.5.

Test statistic

To determine whether global clustering exists, the total number of local clusters is found (for a detailed description on how the local clusters are calculated, see section 2.5). The population size in this potential cluster is compared with the population expected on average in Wales. The expected number of clusters is found by including cases in each window for all geographical units until the p-value exceeds the significance level i.e. there is no longer a significant cluster, if a cluster did exist. This p-value is then multiplied by the number of geographical units. All p-values (multiplied by the number of geographical units) from the "clusters" are summed to create the expected number which is approximately equal to the average of the Monte Carlo distribution. However, if the centroid of a ward is not a cluster or not a significant cluster then these are not included in the overall expected number. For example, if the last significant p-value for every ward was 0.05 then the expected number of clusters would be equal to 45.4 (0.05*908).

Monte Carlo randomisation techniques are used to evaluate the global test statistic by randomising the cases located in the clusters found in the local analysis. The cases are randomly distributed among the population at risk using a multinomial distribution estimated from relative, region specific population sizes. The multinomial distribution is used to distribute cases at random among spatially referenced subgroups where the probability of a case being placed in a particular subgroup is proportional to the population-at-risk size in that subgroup. For example, consider three wards of population sizes 5000 (ward A), 3000 (ward B) and 2000 (ward C). A random number generator supplies values between 0 and 1 such that the value goes into one of the three wards (A, B and C) and counts as a case based on the proportional size of the populations. Therefore any value between 0 and 0.5 is placed in ward A, any value between 0.5 and 0.8 is placed in ward B and any value between 0.8 and 1.0 is placed in ward C.

Results

Ho: The cases are not globally clustered in the dataset.
Ha: The cases are globally clustered in the dataset.

Table 2.11 displays the results for this method with the same leukaemia dataset as previously used. The number of Monte Carlo randomisation runs used was 999 (default value) and the maximum cluster size was arbitrarily chosen as 30 which equates to between 13 and 18 wards in a potential cluster depending on the period of diagnosis examined.

	1982-1986	1987-1991	1992-1996	1997-2001
Cases	1475	1811	2150	2175
Total person years at risk	14175705	14175705	14175705	14175705
Rate per 100000 population	10.405	12.775	15.167	15.343
Obs clusters	65	55	72	93
Exp clusters	29.594	27.769	24.602	24.370
p-value	0.015	0.029	0.001	0.001

Table 2.11: Leukaemia analysis using Besag and Newell's method.

Table 2.11 shows 1811 observed cases of leukaemia for the period 1987-1991 – a rate of 12.775 per 100,000 population. There were 55 observed clusters whereas only 27.769 were expected. This result is significant at the 5% level, p=0.029. All periods are significant indicating that there is significant evidence of global clustering in all datasets.

The dataset of all leukaemia cases in Wales for the period 1982-2001 used in the initial results was analysed again but with 9999 Monte Carlo randomisation runs instead of 999 Monte Carlo randomisation runs to investigate the stability of p-values obtained.

As before, a maximum cluster size of 30 was used for this method with 9999 Monte Carlo randomisation runs and 50 test runs. Table 2.12 is very similar to table 2.11 apart from the range of p-values obtained for the sensitivity analysis. The number of observed clusters and expected clusters show that there were 2.2 times as many clusters as would

have been expected in the period 1982-1986 rising to 3.8 times as many for the period 1997-2001.

	1982-1986	1987-1991	1992-1996	1997-2001
Number of cases	1475	1811	2150	2175
Rate per 100000 population	10.405	12.775	15.167	15.343
Observed clusters	65	55	72	93
Expected clusters	29.594	27.769	24.602	24.37
Mean p-value	0.0172	0.0299	0.0009	0.0001
(Min, Max) p-value	(0.0141,0.0205)	(0.0248,0.0338)	(0.0003,0.0016)	(0.0001,0.0003)

Table 2.12: Summary of leukaemia in Wales 1982-2001, Besag and Newell's method.

For the period 1982-1986, the mean p-value obtained was 0.0172, with values ranging from 0.0141 to 0.0205 indicating that there was evidence of significant clustering in the dataset at the 5% level of significance. For the period 1987-1991, a mean p-value of 0.0299 was obtained with a range from 0.0248 to 0.0338 based on 50 test runs. The period 1992-1996 produced a mean p-value of 0.0009 with a range from 0.0003 to 0.0016 based on 50 runs. The period 1997-2001 produced a mean p-value of 0.0001 with a range from 0.0001 to 0.0003 based on 50 runs. As with all other periods, this showed significant evidence of clustering in the dataset. It can be seen that the range was small for all periods and was very small for the period 1997-2001. For all test runs, the same decision would have been made regarding the significance of clustering in the datasets.

Table 2.13 shows a summary of Besag and Newell's method when the maximum cluster sizes varied between 25 and 35 for two periods, 1987-1991 (the 'least' significant period) and 1997-2001 (the 'most' significant period). These sizes were chosen since 30 was used in the initial analysis. All cluster sizes for the period 1997-2001 produced significant results; however for the period 1987-1991, p-values vary between 0.0109 and 0.1570 depending on the choice of the maximum cluster size indicating unstable results. A maximum cluster size of 32 produces a significant result, p=0.0109. However by increasing the maximum cluster size to 33, an increase of only one case, a non-significant result is obtained p=0.1115. For the period 1987-1991, the number of observed clusters decreases substantially by 22 global clusters, from 66 with a maximum size of 32 and 44

clusters with a maximum size of 33. This was due to many of the global clusters merging into other global clusters since the clusters were very near to each other.

max cluster size	1987-1991				1997-2001			
	obs	exp	obs/exp	p-value	obs	exp	obs/exp	p-value
25	38	26.5475	1.431	0.1532	79	24.8708	3.176	0.0001
26	36	25.4020	1.417	0.1570	95	24.5695	3.867	0.0001
27	43	26.6971	1.611	0.1055	88	24.5369	3.586	0.0001
28	44	26.1529	1.682	0.0731	96	25.2217	3.806	0.0001
29	53	27.5618	1.923	0.0300	100	25.6196	3.903	0.0001
30	55	27.7685	1.981	0.0290	93	24.3696	3.816	0.0001
31	65	28.7216	2.263	0.0165	96	24.7801	3.874	0.0001
32	66	27.8663	2.368	0.0109	89	25.4715	3.494	0.0001
33	44	27.6710	1.590	0.1115	99	25.2202	3.925	0.0001
34	52	28.1248	1.849	0.0662	102	26.5682	3.839	0.0001
35	51	28.1730	1.810	0.0664	103	27.0332	3.810	0.0001

Table 2.13: Sensitivity Analysis for Besag and Newell's method.

The significance of the results changed considerably for each maximum cluster size for the period 1987-1991. Also, many of the global clusters overlapped each other. Hence this method should be disregarded over other methods that are available due to the instability of the method. Another disadvantage of using this method was that you could not control for confounding due to sex and age; this could have made a difference since it is well known that an older population lives along the North Wales coastline. If a large number of clusters were located in these areas then by taking age into account these clusters may have disappeared.

2.4.4. Cuzick and Edwards' Method

The Cuzick and Edwards' method (1990) analyses case point data and control point data at individual level where the control dataset is used to reflect the geographic variation in population density in Wales. This method determines the extent to which nearest neighbours in space are also cases (or controls) for individuals and also whether there are more cases than expected in the k-locations nearest each case. The maximum number of nearest neighbours being cases is user defined. If the kth nearest neighbour is a control then this is not included in the analysis. A nearest neighbour of size 1 from case i is the

closest case to case *i*. A nearest neighbour of size 2 from case *i* is the second closest case to case *i*. Thus a nearest neighbour of size *k* from case *i* is the *k*th closest case to case *i*. The control datasets generated from the NHSAR extract and the lower body cancers diagnosed for the period 1982-2001 were used for this analysis. This method allowed for confounding in the fact that the controls could be matched to the cases by age, sex, deprivation status or any other variables.

Test statistic

The test statistic *T(k)* counts the number of cases that neighbour other cases within *k* nearest neighbours. These counts are summed to make one test statistic *T(k)* for each value of *k*. An adjustment is made to the overall p-value using the Bonferroni and Simes adjustment to take into account the multiple nearest neighbour tests. The test statistic is large when the nearest neighbour to each case is another case. The test statistic is shown in equation 2.5 along with the expected value of the test statistic. The test statistic is summed over all cases. *k* is the user defined maximum number of nearest neighbours to use in the analysis, δ_i is 1 if observation *i* is a case and 0 if observation *i* is a control. d_i^k is equal to 1 if the *k*th nearest neighbour to *i* is a case and 0 if the *k*th nearest neighbour to *i* is a control. c_1 is the total number of cases and c_2 is the total number of cases and controls. The variance is a complex expression and is given by Cuzick and Edwards' (1990).

$$T_k = \sum_{i=1}^{c_1} \delta_i d_i^{\,k}$$

$$E(T_k) = \frac{kc_1(c_1 - 1)}{c_2 - 1}$$

Equation 2.5: Test statistic of Cuzick and Edwards' method.

Results

Ho: The cases are not spatially clustered relative to the controls.

Ha: The cases are spatially clustered compared with the controls.

The leukaemia datasets containing the four time periods were analysed using this method with the maximum number of nearest neighbours k arbitrarily set to 10. The cases along with the NHSAR controls (2 controls per case) and lower body cancers were used for this method. Thus, for each case i, the analysis examines its nearest ten neighbours. For all four periods, significant results were obtained when applying 999 Monte Carlo randomisation runs. Table 2.14 shows the results for the leukaemia dataset for the period 1987-1991.

k	T(k)	E(T)	Var(T)	z	z upper tail p-value	Monte Carlo p-value
1	664	603.444	506.285	2.691	0.004	0.003
2	1321	1206.890	1071.640	3.486	0.000	0.008
3	1945	1810.330	1665.420	3.300	0.000	0.182
4	2555	2413.780	2294.110	2.948	0.002	0.429
5	3186	3017.220	2959.760	3.102	0.001	0.115
6	3824	3620.670	3643.530	3.369	0.000	0.067
7	4461	4224.110	4364.170	3.586	0.000	0.081
8	5105	4827.560	5106.270	3.883	0.000	0.044
9	5749	5431.000	5882.280	4.146	0.000	0.048
10	6361	6034.440	6733.310	3.980	0.000	0.373

Table 2.14: Cuzick and Edwards' analysis for leukaemia dataset 1987-1991 and using NHSAR controls.

T(k) represents the observed numbers of case pairs that were nearest neighbours, *E(T)* represents the expected number of case pairs that were expected to be nearest neighbours, *Var(T)* represents the variance of the test statistic. This is a complex expression and is not shown here but can be viewed in the description by Cuzick and Edwards' (1990). The z score based on the normal distribution is shown along with the corresponding upper tail p-value. A p-value based on the Monte Carlo randomisations is also shown (this is done by shuffling case control labels of the dataset). For $k=4$ and $k=10$, the Monte Carlo p-values are very high compared to other p-values. This is probably due to the randomisation of the original dataset that is used to calculate the p-value (i.e. case and control labels are randomly shuffled) but should be explored further. To allow for

multiple testing, combined p-values are obtained in ClusterSeer V2.2.4 using Bonferroni and Simes adjustments and are shown in table 2.15. Combined p-values are used since for any $k>1$, a statistic is found for each value of k at the same significance level (multiple testing) so ClusterSeer V2.2.4 calculates a combined p-value for all tests performed at one significance level.

	Normal	Monte Carlo
Bonferroni	0.000169	0.030
Simes	0.000034	0.030

Table 2.15: Combined p-values for leukaemia dataset 1987-1991.

Comparing table 2.15 to the nominal 0.05 level of significance, the results for each value of k are now not significant for $k=8$ and $k=9$ using the Monte Carlo results.

Note that this method does not locate the clusters; all that is known is that there is significant clustering in each of the datasets.

Table 2.16 shows an overall summary of the p-values obtained when examining global methods. Figures highlighted in red are significant at the 5% level. Thus all methods gave significant results for the period 1992-1996. Moran's I statistic produces different results from the other methods. This is probably due to rates being used rather than case and population at risk data. Cuzick and Edwards' method produced different results, probably due to the method analysing point data rather than areal data.

P-values	1982-1986	1987-1991	1992-1996	1997-2001
Moran's I Statistic (rook)	0.744	0.674	0.020	0.488
Moran's I Statistic (queen)	0.908	0.606	0.016	0.588
Oden's I Pop	0.014	0.357	0.001	0.018
Besag and Newell	0.015	0.029	0.001	0.001
Cuzick and Edwards	0.001	0.001	0.001	0.001

Table 2.16: Summary of p-values obtained for global methods.

2.5. Local Clustering Methods

Global clustering methods determine whether clustering is present on a global scale i.e. clustering is present in Wales but no specific areas are identified. Local clustering methods determine the specific areas where clusters exist in Wales. i.e. where a much larger number of cases are observed than are expected.

2.5.1. Besag and Newell's Method

Besag and Newell's method has been described previously as a global method. It is also a local method. This method uses case data and population data at grouped (ward) level. The following analysis examines the local method (identifies the specific clusters in Wales). It tries each ward as the centre of a possible cluster. A circular window is expanded to the next ward centroid and is repeated until the specified maximum number of cases in a possible cluster has been identified.

Test statistic

Under the null hypothesis, the population at risk within the circular window is proportional to the case count with a common disease rate throughout the study area. Equation 2.6 shows the calculation of the probability that n_k (the number of wards for each centred ward to contain the maximum cluster size, k cases) has reached or exceeded that predicted by the null hypothesis, N. λ is the average disease frequency multiplied by the population at risk.

$$P(n_k \geq N) = 1 - \sum_{x=0}^{k-1} \frac{e^{-\lambda} \lambda^x}{x!}$$

Equation 2.6: The probability that a local cluster is equal to or greater than Ho.

Equation 2.6 calculates the probability that n_k has reached or exceeded that predicted by the null hypothesis, N (the probability that there are fewer than k cases in the area).

ClusterSeer V2.2.4 lists all clusters with a probability less than the significance level quoted, p=0.05.

Results

Ho: There is no evidence of local clustering in the dataset.
Ha: There is local clustering in the dataset.

The local method produces plots of all significant local clusters along with their respective p-values (not presented here due to the amount of pages of information that are given) and the number of wards aggregated to create the cluster. The locations of the significant local clusters in Wales are shown in figure 2.4 for each period.

For the period 1982-1986, there were three main areas of clustering. Most of the clusters were situated along the North Wales coast. The other areas were Swansea/Carmarthenshire and Blaenau Gwent. Appendix B identifies each of the Local Health Boards in Wales. The clusters are geographically small in size due to these areas being highly populated and hence wards tend to be geographically smaller in size compared to areas such as Mid Wales. For the period 1987-1991, there were more areas of clustering, these being in North West Wales, along the North Wales coast, Mid Wales, West Wales and South East Wales. The geographical size of the clusters is large in Mid Wales due to the wards being geographically larger compared to other parts of Wales due to these areas being less populated. The period 1992-1996 show geographically large clusters along the West Wales coastline and an area of clusters in North Wales. The clusters tend to be scattered throughout Wales for the period 1997-2001 but are mostly concentrated in the Mid Wales area where there are small populations, hence clusters appear to be much larger in geographical size compared to geographically small sized clusters with a much larger population.

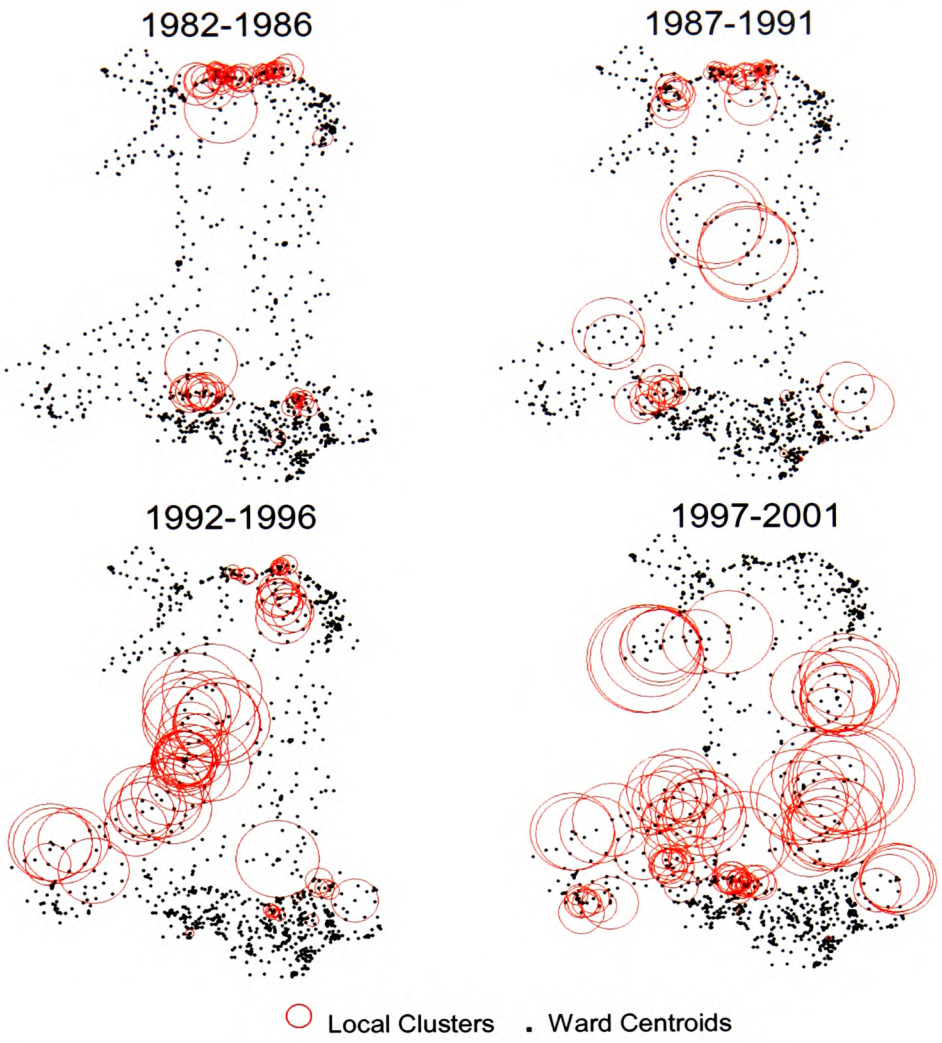


Figure 2.4: Location of clusters in Wales using the leukaemia dataset.

2.5.2. Turnbull's Method

The method of Turnbull et al (1990) is very similar to the previous method. Case data and population at risk data are required at ward level. A predetermined maximum population size p^* (population years) is input into the algorithm. A circular window is centred on each ward and expanded to include adjacent wards until the predetermined population size is finally reached. Note that the furthest region included in the window may only include a fraction of its total ward population and a fraction of the number of cases in that particular ward. Thus, if the population has reached 99,500 and the user inputs 100,000 as the maximum, then the next ward (which has a population of 5,000)

will contribute 10% of its cases and population to the possible cluster. This is repeated for all wards in Wales.

Test statistic

The test statistic is the maximum number of cases amongst all windows of size p^*. There could be many clusters found in the analysis but only the three "most significant" clusters are reported in ClusterSeer V2.2.4.

The significance of the test statistic is calculated via Monte Carlo randomisation (see 2.4.3 for detailed description of this).

Results

Ho: There is no local clustering in the dataset.
Ha: There is local clustering in the dataset.

Table 2.17 summarises the results using Turnbull's method for the leukaemia datasets.

	1982-1986	1987-1991	1992-1996	1997-2001
population p*	288319	234827	197801	195526
cases	1475	1811	2150	2175
total person years at risk	14175705	14175705	14175705	14175705
rate per 100000 population	10.405	12.775	15.167	15.343
most likely cluster p-value	0.001	0.045	0.055	0.054
second most likely cluster p-value	0.002	0.056	0.061	0.056
third most likely cluster p-value	0.002	0.107	0.113	0.127

Table 2.17: Results using Turnbull's method.

The population p^* was chosen so that these results could be compared with the previous method. i.e. the population p^* was calculated to enable the same rate in a cluster as when 30 observed cases was input using Besag and Newell's method to enable a direct comparison. (For example, for the period 1982-1986 there were 1475 cases in a

population at risk of 14,175,705 person-years; thus 30 cases would represent a population at risk of 288,319 person-years). The rate for the period 1997-2001 was 15.343 per 100,000 population. The three most likely clusters for this period were non-significant (the first two clusters are of borderline significance). Three significant clusters were found in the first period 1982-1986 whereas only the most likely cluster was significant for the period 1987-1991. No other results were significant based on a 5% level of significance.

Figure 2.5 shows the location of these clusters and figure 2.6 shows the histograms and test statistic values for the different periods supporting the evidence provided in table 2.17 (thick red line represents the most likely cluster). The histograms estimate the sampling distribution of the test statistic.

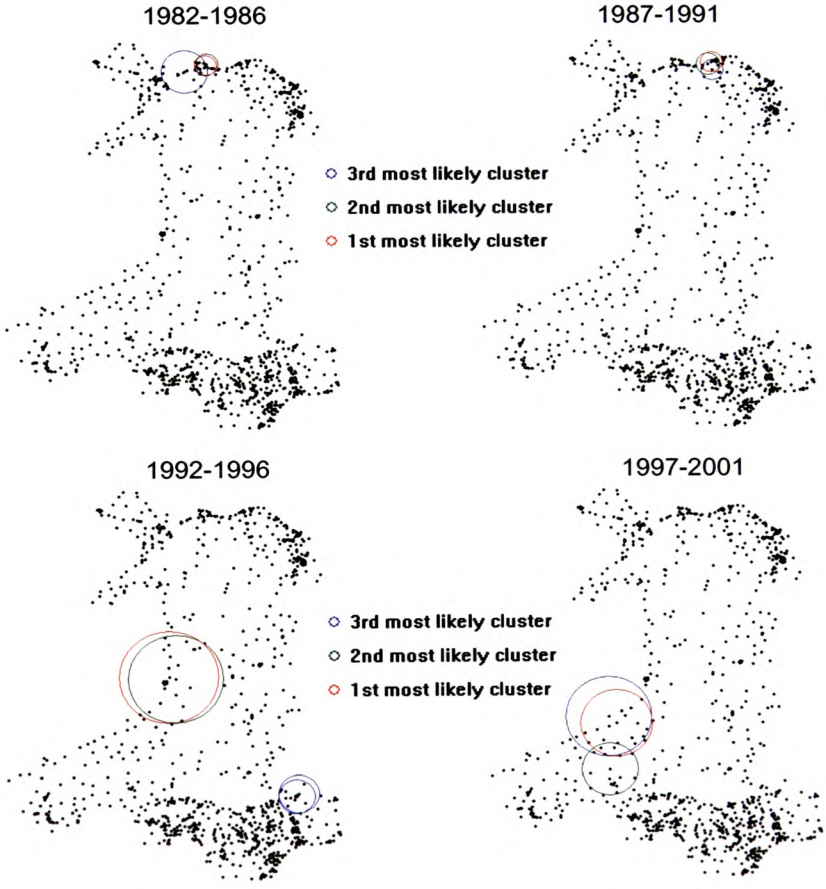


Figure 2.5: Location of most likely clusters using Turnbull's method.

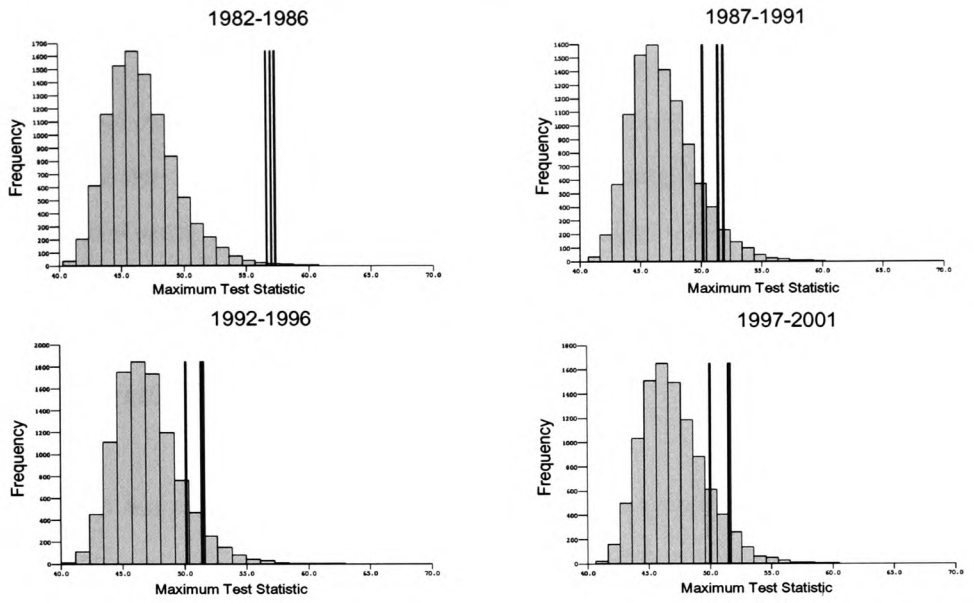


Figure 2.6: Histograms using Turnbull's method.

2.5.3. Kulldorff's Spatial Scan Statistic

The software SaTScan V5.1.3 (Kulldorff, 1997, 2004) was used since the same method in ClusterSeer V2.2.4 could not analyse the dataset. This was due to the amount of memory required to run the dataset. When run in ClusterSeer V2.2.4, an error message would flash on screen during the analysis explaining that the analysis had to stop due to no memory. The method allows the user to work with one of two models. This method uses aggregated case and population at risk data (Poisson model) or case and control data (Bernoulli model).

Test statistic

A circular window is centred on each ward centroid for every ward in Wales and is expanded to include centroids of neighbouring wards. The number of observed and expected cases is noted in each window along with the likelihood of each possible cluster. The likelihood functions for a specific sized window for the Poisson and Bernoulli models are shown in equation 2.7 where c_I is the number of observed cases within the window, C_I is the total number of cases and $E(c_I)$ is the covariate adjusted

expected number of cases within the window under Ho (which is calculated using the population at risk data and the rate in the entire study region) under the Poisson assumption. Under the Bernoulli assumption, c_1 is the number of observed cases within the window, C_1 is the total number of cases, c_2 is the total number of cases and controls within the window and C_2 is the total number of cases and controls. When SaTScan V5.1.3 searches for high rate clusters, x is equal to 1 when the window has more cases than expected under the null-hypothesis and 0 otherwise. The opposite is true when searching for low rate clusters.

Under the Poisson assumption the likelihood function is proportional to

$$x\left(\frac{c_1}{E(c_1)}\right)^{c_1}\left(\frac{C_1-c_1}{C_1-E(c_1)}\right)^{C_1-c_1}$$

Under the Bernoulli assumption the likelihood function is proportional to

$$x\left(\frac{c_1}{c_2}\right)^{c_1}\left(\frac{c_2-c_1}{c_2}\right)^{c_2-c_1}\left(\frac{C_1-c_1}{C_2-c_2}\right)^{C_1-c_1}\left(\frac{(C_2-c_2)-(C_1-c_1)}{C_2-c_2}\right)^{(C_2-c_2)-(C_1-c_1)}$$

Equation 2.7: The likelihood functions under the Poisson and Bernoulli assumptions.

Thus, when case and population data are available, the Poisson model is used whereas when case and control data are available, the Bernoulli model is used.

The likelihood function is maximised over all windows (all possible radii) and a p-value is obtained by repeating the exercise on a large number of random replications in a Monte Carlo simulation. The window that has the maximum likelihood identifies the most likely cluster. The output produces a most likely cluster (if one exists) along with a number of secondary clusters (if any exist). The secondary clusters are other clusters that are ordered by their corresponding likelihoods.

This section uses Kulldorff's spatial scan statistic as a local spatial method using case data and population at risk data and also using case and control data. For the analysis, no clusters overlapped other clusters (the user is able to choose an option as to whether they would like clusters to overlap or not) and the maximum population size as a percentage of the total population at risk was 5%. The default value is 50%. However, this means that a "cluster" could potentially cover over half of the entire country. Thus the value 5% was used. All clusters found were for high rates only.

Results

Ho: The disease does not cluster.
Ha: The disease does cluster.

Leukaemia cases in Wales for the four five year periods covering 1982-2001 and population at risk data were initially analysed as a local method. Table 2.18 shows these results when using case and population at risk data (Poisson model).

	1982-1986	1987-1991	1992-1996	1997-2001
Number of wards	11	2	4	20
Centroid of most likely cluster	306468, 382168	294681, 378151	315397, 233064	243235, 256058
Radius (km)	4.820	1.210	6.290	19.240
Observed Cases in most likely cluster	45	18	14	45
Expected Cases in most likely cluster	19.490	6.190	3.800	24.230
Annual Cases per 100000 population	24.000	37.200	55.700	28.500
Observed/Expected	2.309	2.906	3.683	1.857
p-value	0.001	0.168	0.090	0.190
p-value (secondary 1)	0.011	0.884	0.207	0.209
p-value (secondary 2)	0.708	0.938	0.347	0.256

Table 2.18: Case and population at risk data analysis.

Table 2.18 shows the number of wards in the most likely cluster along with the centroid of the cluster. The radius of the cluster and the number of observed and expected cases were also noted. The rate per 100,000 population found inside the cluster was stated, along with the ratio of observed to expected cases and corresponding p-value. Secondary

clusters show the same information but table 2.18 only shows the p-values for these. A significant cluster was found for the period 1982-1986, p=0.001 of radius 4.8km. The expected number of cases located inside this cluster was 19.490. The Kulldorff statistic was thus ranked highest of all 1000 Monte Carlo randomisations (1/1000=> p=0.001). Table 2.19 summarises the corresponding analysis when using case data with lower body cancers as controls and case data with NHSAR controls.

	Lower Body Cancers				NHSAR Controls			
	1982-1986	**1987-1991**	**1992-1996**	**1997-2001**	**1982-1986**	**1987-1991**	**1992-1996**	**1997-2001**
Number of wards	2	1	32	22	4	34	7	54
Cluster centroid	337795, 186881	257503, 372027	304567, 207883	331917, 204649	321931, 180647	304723, 365836	304323, 199004	186635, 228927
Radius (km)	2.990	0.000	11.300	9.020	1.740	15.980	3.830	35.820
Observed Cases	14	4	131	68	25	82	47	110
Expected Cases	6.160	0.083	95.820	42.750	13.330	58.330	29.330	70.670
Observed/Expected	2.274	4.804	1.367	1.591	1.875	1.406	1.602	1.557
p-value	0.895	0.609	0.181	0.152	0.277	0.317	0.227	0.001
p-value (secondary 1)	0.995	0.690	0.400	0.159	0.281	0.403	0.251	0.037
p-value (secondary 2)	0.998	0.862	0.447	0.400	0.403	0.449	0.505	0.484

Table 2.19: Analysis using case and control data.

There are large differences between results for the case and population at risk data compared with the case and control data. There are also large differences in the p-values for the two control datasets. This may suggest that the control data were not suitable to compare with the cases. For example, for the period 1982-1986, a significant cluster was found in North East Wales using the case and population at risk data. However, non significant clusters were found when using both sets of control datasets with the case data, both in South East Wales. The case and population data entered into SaTScan V5.1.3 did not take age and sex into account (only totals by ward were used) to be consistent with comparable methods in ClusterSeer V2.2.4. This could be a reason as to why the results were so different. Also, it could be that the control datasets are not suitable to use with the case data. Note that if the case and population at risk data are entered into SaTScan V5.1.3 by sex and five year age band, then there are no significant clusters in the dataset.

The locations of the most likely clusters in Wales in table 2.18 and table 2.19 can be found in figure 2.7 along with the locations of the clusters using case and control data.

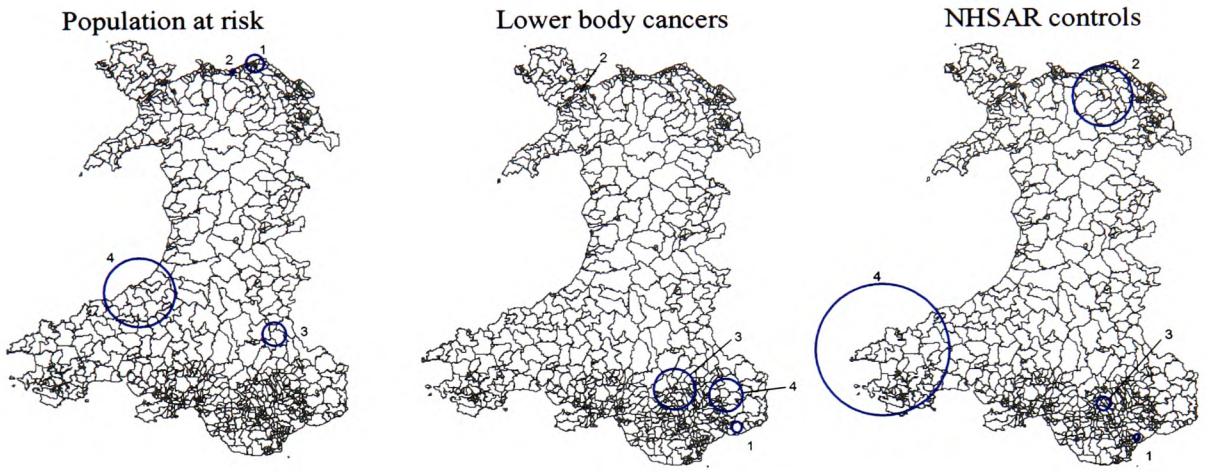


Figure 2.7: Location of most likely clusters for the four periods using the spatial scan statistic.

From figure 2.7 and table 2.18 note the varying results. The NHSAR control data were matched by five year age band, sex and deprivation with the cases. The lower body cancers also have a similar age structure to leukaemia for ages over 25 years. The only significant result, p=0.001, was for the period 1997-2001 using the NHSAR controls; a cluster was located in West Wales. Using the population data and case data the most likely cluster was located in a neighbouring area but was non-significant, p=0.190. The seven wards located as a cluster for the period 1992-1996 using the NHSAR controls were located in the corresponding cluster when using the lower body cancers but both were non-significant.

This method can be used to analyse spatial clustering, temporal clustering or space-time clustering. For this particular analysis temporal clustering has not been analysed. Space-

time clustering is examined in section 2.7. This method will locate low rates as well as high rates in the study region. This method has the advantage over other methods in that there are a number of options that the user can choose for the analysis such as the maximum cluster size (as a percentage of the total population at risk). Another option is for no geographic overlap – this is the default and enables no clusters to overlap each other. The user guide states that p-values for secondary clusters are considered as conservative due to the method in which the likelihood ratio test only tests the null hypothesis against the alternative hypothesis that there is only one cluster.

2.5.3.1. Turnbull v Kulldorff

Turnbull's method and Kulldorff's method both require the user to input a maximum population size in a cluster to be identified. Thus, the results from both these methods can be compared with each other. Results are compared in terms of multiple runs of Turnbull's method and 9999 Monte Carlo simulations to determine the validity of results.

Kulldorff's Scan Statistic for SaTScan V5.1.3 and ClusterSeer V2.2.4 use pseudo-random number generators and are used for random replications of the data set under the null hypothesis. Using SaTScan V5.1.3, the same p-value was obtained if the same dataset was rerun. This is due to the pseudo-random number generator; the seed is set as the same value for each run and is not changeable in SaTScan V5.1.3, hence if the input data are the same then so will be the output. This is not the case in ClusterSeer V2.2.4. All runs will produce a different seed and hence, different p-values in ClusterSeer V2.2.4.

The Kulldorff Scan Statistic was compared with Turnbull's method as a local clustering method. Both methods analyse aggregated data at ward level and take into account the population at risk. Kulldorff's method does not run in ClusterSeer V2.2.4 due to the large dataset used and memory required, hence the software SaTScan V5.1.3 was used. Note that the p-value obtained using Kulldorff's method is not compared when repeating the test a number of times due to the pseudo random number generator explained earlier.

For Turnbull's method, the population figures used for each period were the same as used in the initial analysis. Each of the four datasets was rerun 50 times using 9999 Monte Carlo randomisation runs. The first three clusters are produced in the output and ordered by significance. Table 2.20 shows the results of the 50 test runs using Turnbull's method.

As can be seen from table 2.20, the p-values obtained for all tests have small ranges. However the most likely clusters for the time periods 1992-1996 and 1997-2001 show a conflicting decision. For example, for the period 1997-2001, a minimum p-value of 0.0488 indicates a significant result. However the mean p-value and maximum p-value indicate a non-significant result. Care should be taken when one test is run of borderline significance since multiple test runs may show contradictory results. It should be noted that the same cluster is found in each of the periods for all runs.

	1982-1986			1987-1991		
	Cluster 1	**Cluster 2**	**Cluster 3**	**Cluster 1**	**Cluster 2**	**Cluster 3**
Mean p-value	0.0014	0.0018	0.0022	0.0453	0.0556	0.1062
Min p-value	0.0007	0.0011	0.0014	0.0407	0.0526	0.1006
Max p-value	0.0023	0.0030	0.0035	0.0490	0.0599	0.1137

	1992-1996			1997-2001		
	Cluster 1	**Cluster 2**	**Cluster 3**	**Cluster 1**	**Cluster 2**	**Cluster 3**
Mean p-value	0.0535	0.0590	0.1133	0.0528	0.0555	0.1289
Min p-value	0.0483	0.0547	0.1053	0.0488	0.0509	0.1231
Max p-value	0.0605	0.0652	0.1196	0.0569	0.0600	0.1356

Table 2.20: Results using Turnbull's method.

For Kulldorff's spatial scan statistic, the seed is set as the same value for each run. Therefore this method is not run several times; it is only run once to obtain a p-value to 4 decimal places using 9999 randomisation runs. Exploring the algorithms with a higher number of Monte Carlo randomisation runs provides more accurate results i.e. further decimal places. Thus a significant result of 0.001 using 999 MC randomisations may be a highly significant result at 0.00001 if 99,999 MC randomisations were used.

All leukaemia cases in Wales for the four time periods 1982-1986, 1987-1991, 1992-1996 and 1997-2001 were analysed using the Kulldorff Spatial Scan Statistic. To compare these results with Turnbull's method, the maximum cluster size using case and population at risk data were adjusted accordingly to the population figures used for each time period using Turnbull's method. Table 2.21 illustrates this for all time periods used. e.g. For the period 1982-1986, 2.03% corresponds to 288,319 person years for the five year period. A percentage of the total population at risk is used for Kulldorff's method whereas Turnbull's method requires a population figure.

	All Wales	Turnbull	Kulldorff
1982-1986	14175705	288319	2.03%
1987-1991	14175705	234827	1.66%
1992-1996	14175705	197801	1.40%
1997-2001	14175705	195526	1.38%

Table 2.21: Person-years for Turnbull and Kulldorff's methods.

Table 2.22 summarises the results for the four time-periods using the spatial scan statistic by Kulldorff.

	1982-1986	1987-1991	1992-1996	1997-2001
Number of wards	11	2	4	20
Radius	4.8km	1.2km	6.3km	19.2km
Observed cases	45	18	14	45
Expected cases	19.4900	6.1900	3.8000	24.2300
Annual cases per 100000 population	24.0000	37.2000	55.7000	28.5000
Observed/Expected	2.3090	2.9060	3.6830	1.8570
p-value	0.0009	0.1676	0.0899	0.1896

Table 2.22: Summary of most likely clusters in Wales using Kulldorff's method.

Table 2.22 shows that for the period 1997-2001, 45 cases were observed within a radius of 19.2km. 24.2 cases were expected, giving a non-significant p-value of 0.1896 indicating no significant clustering. The first time period 1982-1986 produced the only significant result at 0.0009 of radius 4.8km. Kulldorff's method also produces secondary clusters although these are not shown in table 2.22.

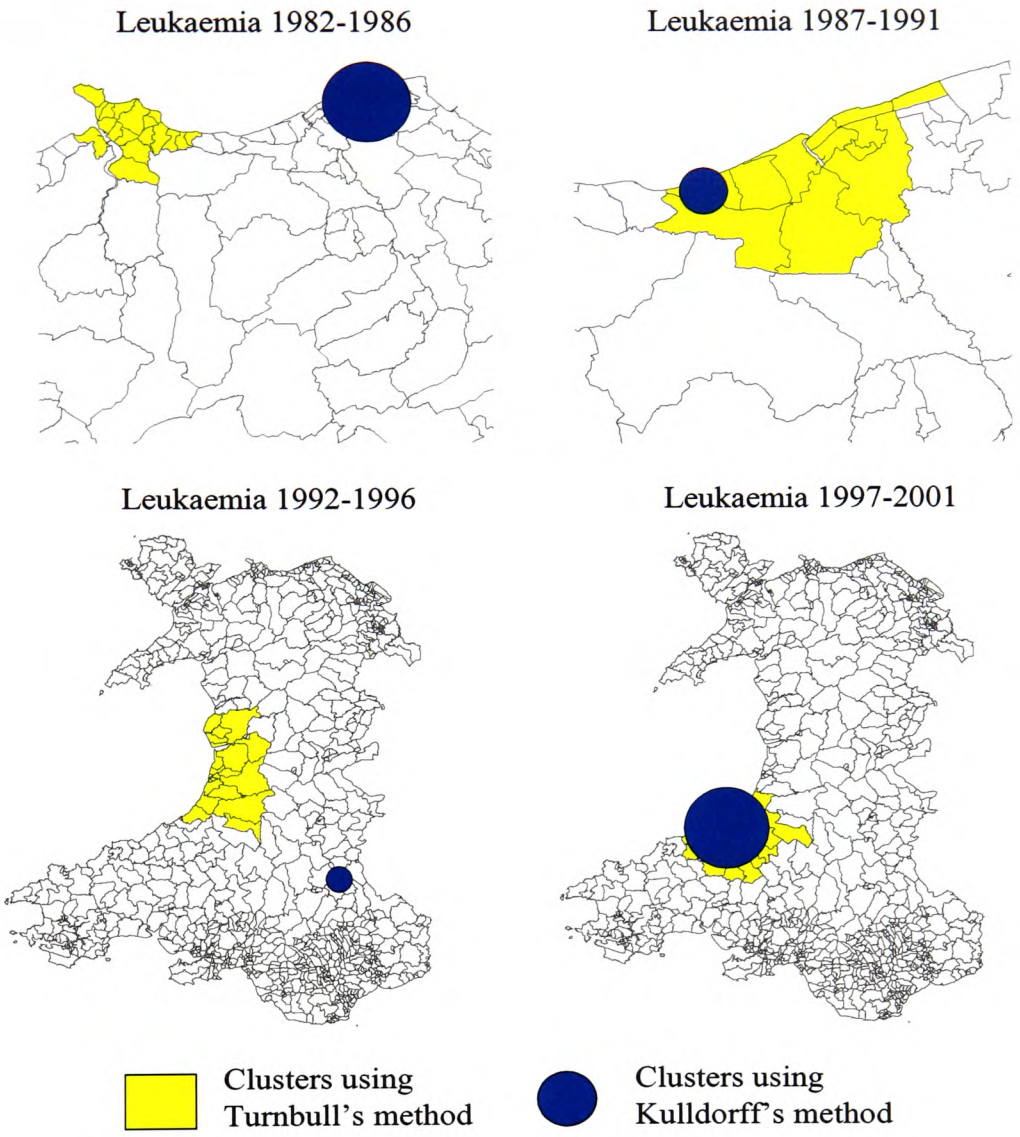


Figure 2.8: Location of clusters using Kulldorff's method and Turnbull's method.

Comparing this method with Turnbull, the secondary clusters from Kulldorff's method cannot be compared due to conservative p-values. However, the advantage of Kulldorff's method is that resulting clusters do not overlap. For Turnbull's method, at least two of the clusters overlap in each time-period analysed. Examining the most likely cluster, the locations of the clusters using Kulldorff's method and Turnbull's method are shown in figure 2.8. For the time periods 1987-1991 and 1997-2001 the wards identified as the most likely cluster in the Kulldorff method are also identified using Turnbull's method -

although Turnbull's method identifies additional wards for these time periods since it always has to identify the maximum population size. For the period 1982-1986 both methods identified a cluster along the North Wales coast, approximately 25km apart. However, for the time period 1992-1996, Turnbull's method identified a cluster on the West Wales coast but Kulldorff's method identified a cluster further south and east on the Powys/Gwent border. Comparing the two methods they do not produce the same sized cluster. The reason for this is that for Turnbull's method the algorithm aggregates wards until it has reached the maximum population level for every ward centroid in Wales. Thus the highest number of cases is always output as the most significant. However, for Kulldorff's method, the located clusters are allowed to vary in population size, thus a cluster may be located with a smaller population level than what was input. Therefore Kulldorff's method has the advantage over Turnbull's method in that it can identify smaller clusters than the maximum population level, whereas Turnbull's method will only identify clusters at the maximum population level. This is why the clusters found using Kulldorff's method are generally smaller than those using Turnbull's method. It is postulated that clusters should be "more" significant using the spatial scan statistic compared with Turnbull's method since the spatial scan statistic can take any radius from 0 to the upper limit whereas for Turnbull's method each potential cluster is only one size with the maximum number of cases. This is due to the fact that the spatial scan statistic takes multiple testing into account and adjusts the p-value obtained. Additionally, from Kulldorff's analysis, it was evident that confounding should be taken into account in which Turnbull's method in ClusterSeer V2.2.4 cannot.

Further work has been carried out using simulated datasets (section 2.9) to identify which method is better at identifying where actual clusters are located.

2.5.4. Anselin's Local Moran Test

Anselin's Local Moran statistic (1995) is a weighted correlation coefficient for aggregated data (at small area level) and can detect significant outliers in datasets and is generally used to detect local spatial clusters. Note an outlier is defined as an areal unit that has a much higher or lower risk or rate than its surrounding areal units. The method

requires case data at aggregated level – population at risk data are not taken into account. The local Moran value test detects local spatial autocorrelation by using local indicators of spatial autocorrelation (LISAs) at each geographical unit. The Local Moran test breaks down Moran's I to ward level analysis termed LISAs which in turn gives an indication of the extent to which spatial clustering of similar values around that specific geographical unit occurs.

Test statistic

LISAs are local statistics that quantify spatial autocorrelation and clustering of either similar (or dissimilar) disease frequency values at small area level. The local Moran test statistic will be positive when values at neighbouring wards are similar and negative otherwise. This method is able to identify significant clusters in the absence of global autocorrelation.

Anselin's local Moran statistic is given by $I_i = d_i \sum_j w_{ij} d_j$ where d_i is the difference between the observed numbers of cases in area i and the mean observed cases for all areas. w_{ij} is a weight based on area i and area j connectivity that enables only neighbouring values of d_j to be included in the test statistic. The weights are standardised to take into account the number of neighbours (e.g. a ward that has three neighbours would give corresponding weights of a 1/3 to each of its neighbours to enable the weights to sum to 1).

In summary, Moran's I is broken down into LISAs for each ward in Wales. The sum of LISAs are proportional to Moran's I value. Thus LISAs can be used to identify outliers in global patterns or as an indicator of a local spatial cluster. Significance is calculated via conditional randomness (see 2.2). The significance level of each area is adjusted to take into account the multiple tests at each ward in Wales. This test also identifies low value outliers as well as high value outliers. All 908 wards in Wales are analysed to determine if they are outliers compared with their neighbouring wards.

Results

Ho: There is no association between case counts in neighbouring areas. I_i is 0.

Ha: $I_i \neq 0$.

Table 2.23 shows the results of the leukaemia dataset analysis using Anselin's local Moran test statistic and 999 Monte Carlo randomisation runs.

	1982-1986	1987-1991	1992-1996	1997-2001
Cases	1475	1811	2150	2175
Average frequency per ward	1.624	1.994	2.368	2.395
No. wards < 0.05	79	58	70	66
No. wards < 0.05 (L, H)	(26, 53)	(21, 37)	(26, 44)	(24, 42)
No. wards < 0.009954	26	16	13	19
No. wards < 0.009954 (L, H)	(8, 18)	(5, 11)	(5, 8)	(8, 11)

Table 2.23: Analysis using Anselin's local Moran method.

As can be seen from table 2.23, there were 1475 leukaemia cases for the period 1982-1986, giving an average leukaemia frequency of 1.624 cases per ward. 79 wards were identified with a p-value less than 0.05 (via Monte Carlo randomisations). 26 of these wards were considered low value outliers (L) and 53 wards were considered as high value outliers (H). Since local Moran statistics tend to be correlated among neighbouring locations, the level of significance was adjusted to allow for several locations being considered simultaneously. There was an average of 5.233 contiguous neighbours for each ward. The lower of the two adjusted measures was Bonferroni with an adjusted significance level of 0.009554. Thus p-values less than 0.009554 are considered significant based on the Bonferroni adjustment. There were 26 wards considered significant by the Bonferroni adjustment of which 8 wards were of low value outliers and 18 wards of high value outliers. Figure 2.9 identifies all significant outliers based on the Bonferroni adjusted significance level for all periods of the leukaemia dataset.

Most outliers tend to be situated in the Southern half of Wales; this is where the highest populations are found. South Wales displays a large number of high value outliers

whereas the low value outliers tend to be in mid-Wales, areas of low population. As with Moran's I statistic this test identifies the number of cases for each ward and examines whether neighbouring wards were similar or not. Large population differences between areas will bias the results. These results should be treated with caution due to the differing population sizes of Welsh wards from the 1991 UK Census. Further analysis is required by looking at the population sizes of these outliers to see if the populations are similar or dissimilar to neighbouring areas.

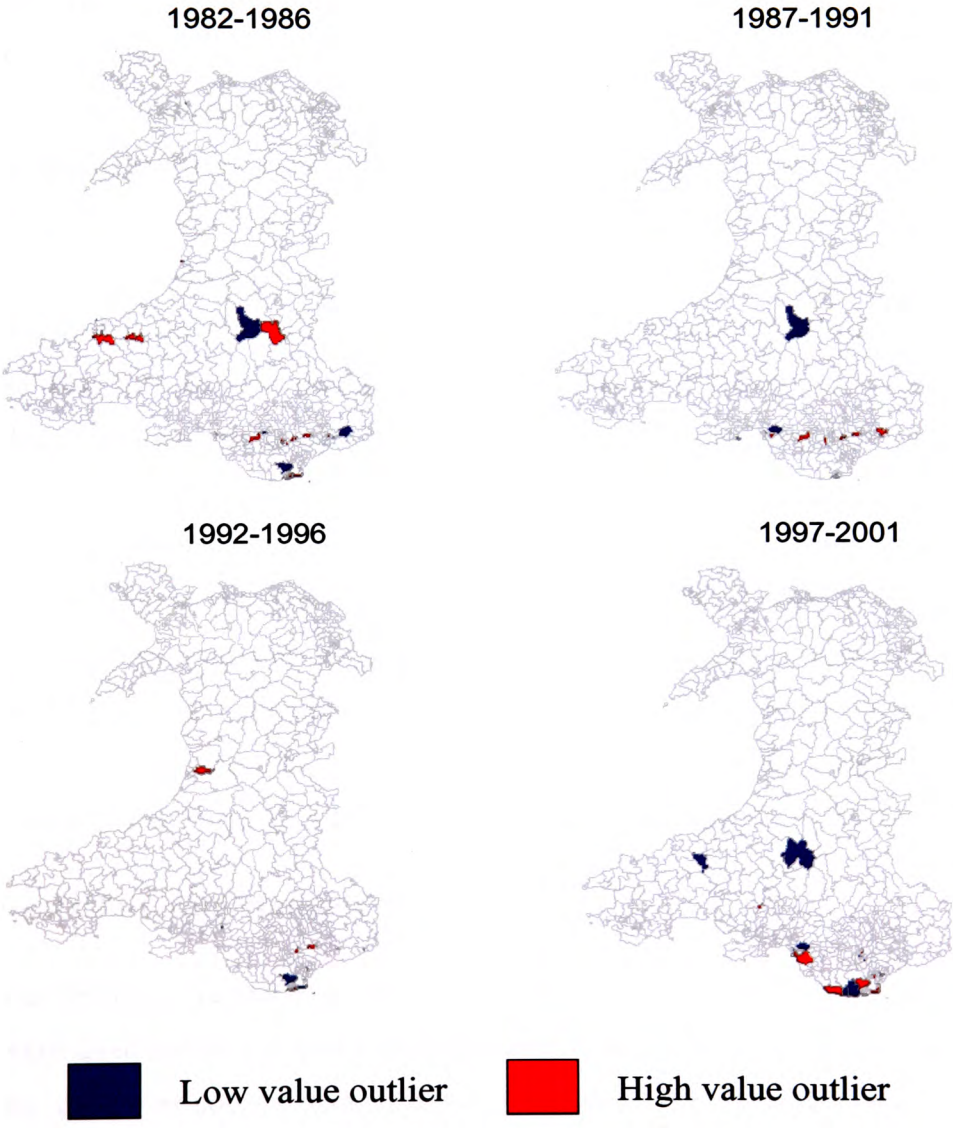


Figure 2.9: High and low value outliers using Anselin's local Moran test.

Summary

To summarise, for the periods 1982-1986 and 1987-1991, Besag and Newell, Turnbull and Kulldorff all show clusters along the North Wales coast when using similar population sizes for clusters. These clusters were not in the same areas along the North Wales coast. Besag and Newell, Turnbull and Kulldorff all agree that most likely clusters are located in West Wales for the period 1997-2001 when similar cluster population sizes are used. However, the period 1992-1996 shows conflicting results. Both Besag and Newell and Turnbull agree that clusters are located in West Wales although Besag and Newell also locates clusters in Mid-Wales. However, Kulldorff's method and Besag and Newell's method identifies clusters in South Powys whereas Turnbull does not. Besag and Newell's method produces a large number of clusters in the analysis and only a map is provided for these clusters – no information is given regarding the number of observed cases or expected cases in the cluster. Thus, other methods shown here provide more information than this so this method should not be used for cluster analysis. From Anselin's local Moran test it can be seen that there are various high and low value outliers in Wales for the varying periods. For the period 1992-1996, a high value outlier was located in West Wales where the cluster using Turnbull's method was located. This may have contributed to the most likely cluster using Turnbull's method being located in this area. Further investigation should be carried out on these outliers to determine the unusual numbers of cases obtained. Thus, Anselin's method can be used to determine unusual numbers of cases in particular areas compared with neighbouring areas. Comparisons between Turnbull's method and Kulldorff's method are discussed later.

2.6. Focused Clustering Methods

In order to investigate the use of the following test statistics, potentially important spatial locations were used based on previous literature. The datasets were analysed around a point source, the point source being Nant-Y-Gwyddon (NYG) landfill site in the South Wales Rhondda Valleys, easting 297975, northing 193986. The WCISU have analysed this landfill site in the past in relation to a possible link with increased incidence of non-Hodgkin's lymphoma (WCISU, 2004). The site opened in 1988 and closed down in 2002

due to public pressure. Therefore the analysis has been split into pre opening of the site (1982-1987) and post opening of the site (1988-2001).

2.6.1. Kulldorff's Spatial Scan Statistic

Ho: The disease does not cluster around the point source.
Ha: The disease does cluster around the point source.

The spatial scan statistic is described in section 2.4.2.3. The local method tests every possible ward centroid in the study region as the centre of a possible cluster. However, for the focused method, a separate file containing one pair of coordinates (an easting and a northing) is input by the user so that only this point source coordinate is used as the possible focus of a cluster. This is the only grid reference(s) that a window will be expanded upon to determine the significance. 999 Monte Carlo randomisation runs were used for the analysis.

Results

Table 2.24 summarises the results around NYG landfill site using case and population at risk data. Note that only data within 20km of NYG landfill site was used for all focused methods.

	1982-1987	1988-2001
Number of wards	4	3
Radius (km)	1.73km	1.53km
Person-years in cluster	94356	170604
Observed Cases	13	24
Expected Cases	8.64	23.05
Annual cases/100000 population	13.8	14.1
Observed/Expected	1.505	1.041
p-value	0.157	0.673

Table 2.24: Case data and population at risk data analysis.

Similar radii were located as the most likely cluster around NYG landfill site for both periods, but both clusters were non-significant. The higher ratio of observed to expected cases was found for pre-opening of the site at 1.505 compared with 1.041 after the site had opened. Table 2.25 shows similar analysis when using the NHSAR and lower body cancer control datasets with the leukaemia dataset.

	NHSAR controls		LBC controls	
	1982-1987	1988-2001	1982-1987	1988-2001
Number of wards	No cluster was identified for this period	No cluster was identified for this period	4	No cluster was identified for this period
Radius (km)			1.73km	
Population (cases and controls)			83	
Observed Cases			13	
Expected Cases			11.28	
Observed/Expected			1.152	
p-value			0.444	

Table 2.25: Case data and control data analysis around NYG landfill site.

Note that no clusters were found in either period examined when using the NHSAR controls. No cluster was identified for the later period 1988-2001 using the lower body cancers either, but a non-significant most likely cluster was found for the earlier period. The difference between finding a non-significant cluster and not finding a cluster at all is that the number of observed cases is greater than the number of expected cases but is not significant (non-significant cluster) whereas not finding a cluster at all means that at no radii were the observed cases greater than the expected cases within the specified distance. This cluster was the same size cluster found when using case and population at risk data. However, although both clusters contained increased risks of leukaemia, both were non-significant. It is not surprising that the case and control data did not find a cluster in the period 1988-2001 since the "cluster" found using the case and population data was only 1 observed case higher than what was expected and this result was highly non-significant.

2.6.2. Score Test of Lawson and Waller

The Score test, proposed by Lawson (1989) and Waller et al (1992), requires case and population at risk data at aggregated level. The Score test is a focused clustering test and

assigns a score to each ward based on the difference between observed and expected counts and weighted by inverse distance to the focus (i.e. the weight is proportional to the inverse of the distance). Thus a case 2km from the source would obtain a weighting of 0.5 and a case 4km from the source would yield a weight of 0.25.

Test statistic

The test statistic for this method is the sum of the differences between the observed counts O_i and the expected counts E_i for all wards n in Wales. These differences are weighted by the inverse distance from the focus to a particular ward centroid $\frac{1}{d_i}$. The test statistic can be viewed in equation 2.8.

$$T = \sum_{i=1}^{n} \frac{(O_i - E_i)}{d_i}$$

Equation 2.8: Test statistic using the Score test of Lawson and Waller.

Under the null hypothesis Ho, the test statistic has mean 0. Equation 2.9 shows the standardised statistic T_N, which has an asymptotic standard normal distribution except for very rare diseases.

$$T_N = \frac{T}{\sqrt{Var(T)}}$$

$$Var(T) \cong \sum_{i=1}^{n} \left(\frac{E_i}{{d_i}^2} - O_i \sum_{i=1}^{n} \frac{p_i}{d_i p*} \right)$$

Equation 2.9: The standardised statistic T_N and the variance of T_N.

p_i denotes the population in area i and $p*$ denotes the total population size.

Results

Ho: The disease does not cluster around the point source.

Ha: The disease does cluster around the point source.

NYG landfill site was selected as the focus. The same time periods as the previous method were used. The datasets used were within 20km of NYG landfill site. Cases and controls in other parts of Wales are nearer to other landfill sites that may influence the risk in those particular areas. Table 2.26 summarises the results.

	1982-1987	1988-2001
Number of wards	130	130
Number of cases	305	1050
Total person years	3331224	7772856
Average disease frequency*	9.15779	13.5085
Test statistic	1.055	-0.582
Normal p-value	0.146	0.720
Monte Carlo p-value	0.201	0.667

* per 100,000 population

Table 2.26: Analysis of leukaemia within 20km of NYG using the Score test.

This method produced non-significant results with Monte Carlo p-values from 0.201 in 1982-1987 to 0.667 in 1988-2001. These p-values were similar to what were obtained using approximations from a standard normal distribution. Figure 2.10 shows the histograms obtained when using this method along with cumulative observed and expected plots, supporting the evidence produced in table 2.26.

Figure 2.10 shows the cumulative plots from NYG landfill site within 20km. The number of observed and expected cases within 20km of the focus is very similar to each other at all radii (i.e. no large deviations). The histograms (using the MC approach) shown in figure 2.10 supports the evidence of results provided here. Kulldorff's method did not identify any clusters around the landfill site during 1988-2001 when using case and population at risk data. The observed-expected figures in figure 2.10 also support Kulldorff's results when using case and population at risk data.

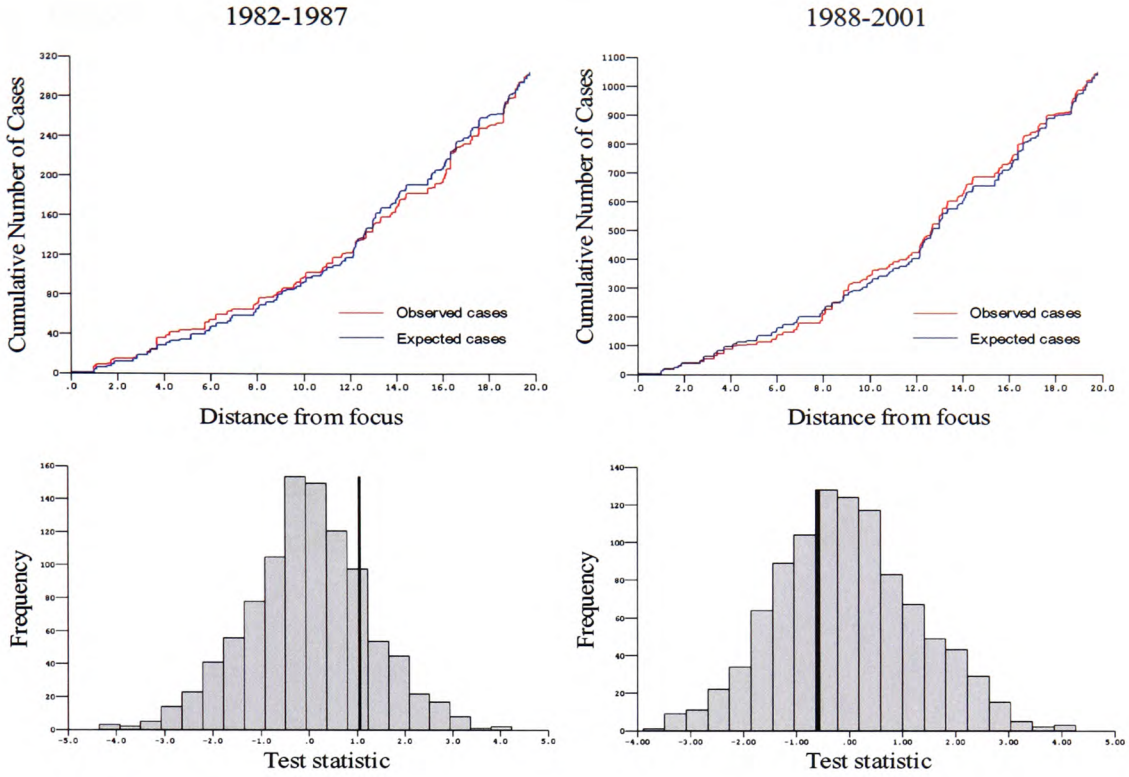


Figure 2.10: Histograms and cumulative plots using the Score test.

The Score test of Lawson and Waller was run a further 50 times, each time using 9999 Monte Carlo randomisation runs for each of the datasets to determine the stability of the p-values obtained. Table 2.27 shows the results for the Score test.

	1982-1987	1988-2001
Mean p-value	0.2028	0.6655
Min p-value	0.1931	0.6557
Max p-value	0.2131	0.6741
Range	0.0200	0.0184

Table 2.27: p-values for multiple run analysis.

Table 2.27 shows no conflicting decisions regarding significance. However, as in previous methods, care must be taken when borderline significant results are obtained. These should be run a number of times to enable a decision regarding a significant or non-significant result.

2.6.3. Bithell's Linear Risk Score Test

Bithell's method (1995, 1999) requires aggregated case and population at risk data at a specific geographical unit. In this case Welsh wards were used from the 1991 UK Census.

Ho: The disease does not cluster around the point source.
Ha: The disease does cluster around the point source.

A parametric risk model is proposed for the alternative hypothesis. One of four models is chosen to test the alternative hypothesis. These "relative risk functions" (RRF) are shown in equation 2.10.

$$RRF_1 = e^{\frac{\phi}{d}}$$

$$RRF_2 = 1 + \beta e^{-\frac{d}{\phi}}$$

$$RRF_3 = 1 + \beta e^{-\left(\frac{d}{\phi}\right)^2}$$

$$RRF_4 = 1 + \frac{\beta}{1 + \frac{d}{\phi}}$$

Equation 2.10: Relative risk functions.

RRF1 can be discarded for this analysis as it is infinite at the focus. Values for $(1+\beta)$ (ratio of risk at focus over that infinitely far from focus) and ϕ (rate of decay of cases with distance from the focus) are input into the model. The RRF to use can be determined from a plot of all four RRFs with the observed data. All relative risk functions tend to 1 as distance increases towards infinity. When the RRF is equal to 1 the risk is the baseline. ϕ determines how quickly the RRF tends to 1. The higher the value of ϕ the slower the RRF gets to 1. The default values in ClusterSeer V2.2.4 are 0 and 0.01 for β and ϕ respectively. Note the value of ϕ to use depends on the unit of distance being used. Under the null hypothesis of no clustering, $\beta=0$ (null hypothesis of no

clustering). Units for *d,* the distance from the focus must be consistent with the units used for the eastings and northings in the dataset.

Test Statistic

The method assigns a risk score to each case, the logarithm of the relative risk for that particular ward. The underlying model for the number of cases in a ward follows a Poisson distribution. A log likelihood test is used to identify whether the null hypothesis or the alternative hypothesis better fits the data. The log likelihood function (LLF) is shown in equation 2.11 where λ_{a_i} is the relative risk under the alternative hypothesis for ward i (from one of the RRFs selected), λ_{0_i} is the relative risk under the null hypothesis for ward i (a constant), an expected count e_i is modified by these relative risks and k is the number of wards and x_i is the number of cases in ward i.

$$LLF = \sum_{i=1}^{k}\left[x_i \log\left(\frac{\lambda_{a_i}}{\lambda_{0_i}}\right) - e_i(\lambda_{a_i} - \lambda_{0_i}) \right]$$

Equation 2.11: Log likelihood function.

The most powerful test of the null versus the alternative hypothesis is whether the test statistic exceeds a critical value c_0, and is chosen based on an appropriate type 1 error (the probability of rejecting the null hypothesis when it is actually true). Thus, from equation 2.11, the test statistic T is shown in equation 2.12 where λ_{a_i} is the relative risk under the alternative hypothesis for ward i, x_i is the number of cases in ward i and n is the total number of wards.

$$T = \sum_{i=1}^{k} x_i \log(\lambda_{a_i})$$

Equation 2.12: Test statistic using Bithell's linear risk score test.

This method allows the user to run an unconditional test or a conditional test. For an unconditional test, the relative risk is constant across regions and equals 1 under the null hypothesis. The baseline disease frequency used to calculate expected case counts for

distance from the focus (and thus the relative risk (the number of observed cases divided by the number of expected cases)) is assumed appropriate for the study area. For a conditional test, the relative risk is assumed to be constant across regions, but not necessarily equal to 1 under the null hypothesis. The baseline disease frequency used to calculate expected case counts is not assumed appropriate for the study area. The unconditional test is used for this analysis since the baseline disease frequency is used to calculate expected counts.

Bithell's test can be interpreted in two ways; hypothesis testing or model fitting. Under hypothesis testing, the test concludes whether there is clustering or not (parameters are chosen objectively) while under model fitting parameter estimation is used to get the best match to the pattern of the data and hence the p-value should not be interpreted as significant clustering (or no significant clustering) since it tests a hypothesis generated for the data using the data. i.e. circular reasoning. Both methods are shown here.

The user enters initial values of β and ϕ into the model and Bithell's method produces optimum values for β and ϕ along with a p-value based on the RRF selected. The p-value is obtained via observations from the original dataset being randomised (following the calculation of the test statistic). The statistic is recalculated and this process is repeated a number of times to amass distributions that are used to calculate the p-value.

Results (Model Fitting)

Using the leukaemia data within 20km of NYG landfill site, cumulative plots of observed and expected numbers of cases and histograms were obtained. Various values of β and ϕ were input into the model and the resulting p-values obtained. Table 2.28 summarises the p-values obtained when varying the values of β and ϕ to model the case and population dataset for the four relative risk functions.

1982-1987						
	RRF1			RRF2		
$\Phi \backslash \beta$	1	2	3	1	2	3
1	0.224	0.217	0.216	0.141	0.147	0.177
2	0.214	0.198	0.221	0.156	0.149	0.172
3	0.223	0.185	0.208	0.209	0.240	0.184
	RRF3			RRF4		
$\Phi \backslash \beta$	1	2	3	1	2	3
1	0.115	0.123	0.128	0.316	0.299	0.340
2	0.181	0.209	0.229	0.364	0.371	0.395
3	0.159	0.172	0.168	0.403	0.406	0.428
1988-2001						
	RRF1			RRF2		
$\Phi \backslash \beta$	1	2	3	1	2	3
1	0.666	0.660	0.679	0.753	0.710	0.750
2	0.668	0.679	0.678	0.828	0.840	0.839
3	0.652	0.642	0.656	0.806	0.751	0.747
	RRF3			RRF4		
$\Phi \backslash \beta$	1	2	3	1	2	3
1	0.589	0.593	0.589	0.634	0.612	0.599
2	0.722	0.727	0.748	0.555	0.539	0.548
3	0.862	0.861	0.873	0.565	0.516	0.499

Table 2.28: Resultant p-values obtained for specified initial values of β and ϕ.

Table 2.28 shows a sample of the results since many initial values of β and ϕ should be used to model the fit of the data to produce the optimum results. The p-values here refer to how "well" the model fits the data as opposed to the significance. For example, the p-values obtained for the period 1988-2001 range from 0.499 (β =3, ϕ=3) using RR4 to 0.873 (β =3, ϕ=3) using RRF3. It should be noted that when data are sparse, the resulting optimum values changed depending on the initial values entered. This was found when analysing data around another landfill site. The models for NYG are shown in figure 2.11 when β=3 and ϕ=3.

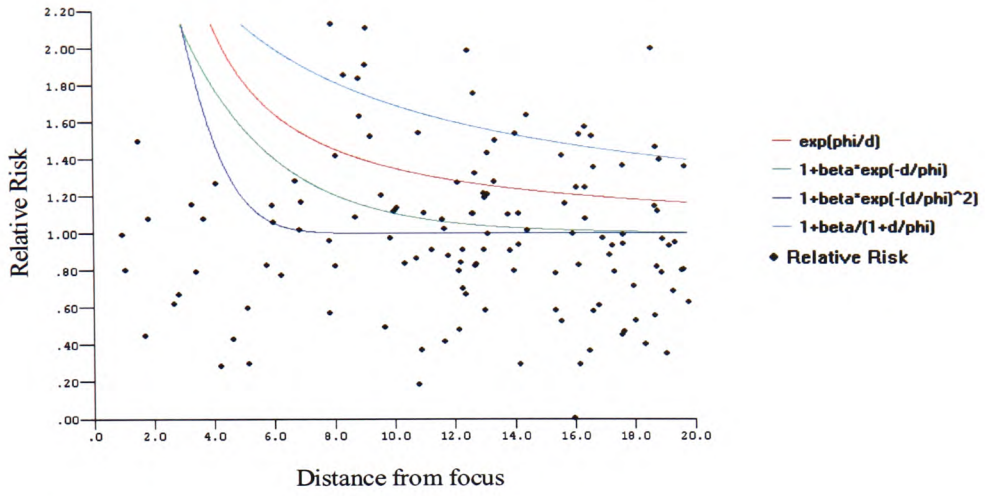


Figure 2.11: The model fit for the period 1988-2001 using original values β=3 and ϕ=3.

Results (Hypothesis Testing)

The same dataset as above was used for the analysis. However, instead of entering various values of β and ϕ , assume that a relative risk of 2 exists at the focus and that the relative risk gradually decreases towards 1 as distance from the focus increases. Figure 2.12 shows the relative risk functions for this situation when β=1 and ϕ=3 along with the histograms for each of the periods when RRF4 was used. Table 2.29 shows the resulting analysis when using RRF4.

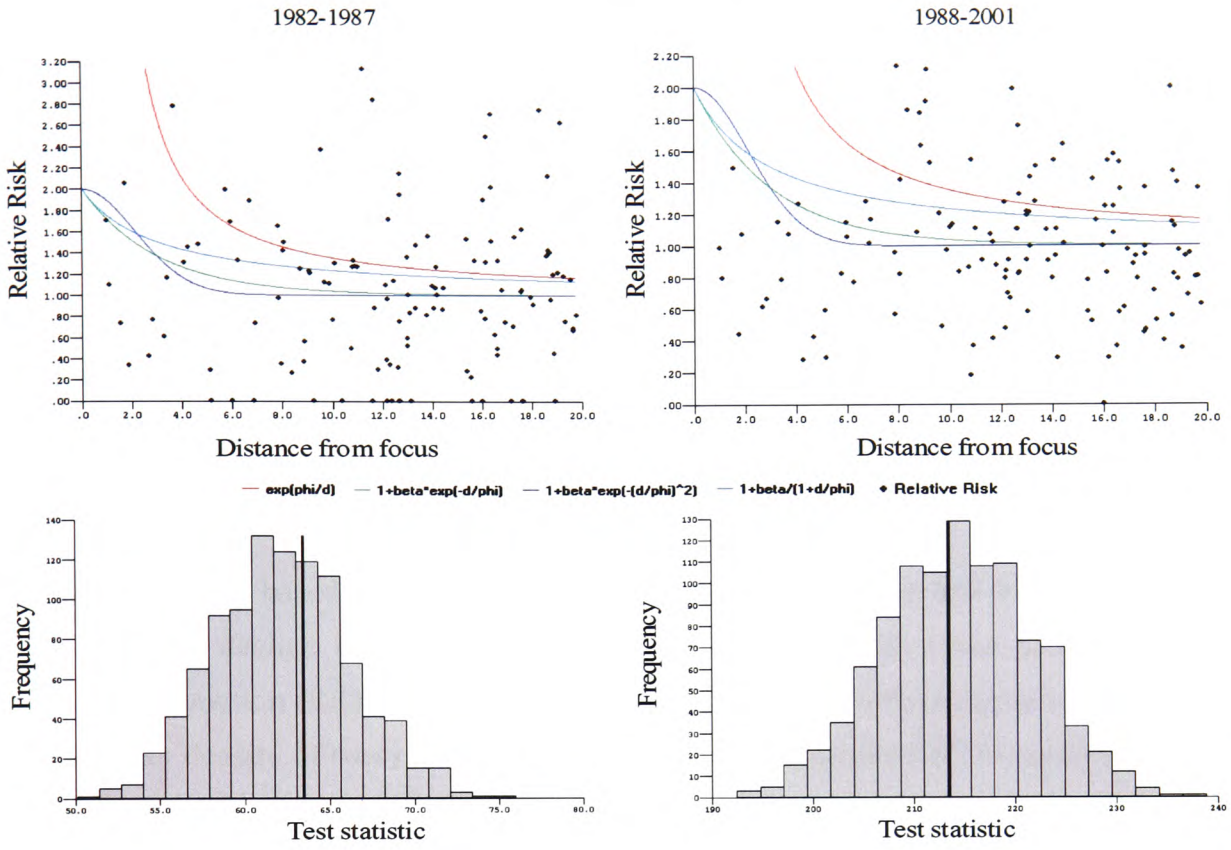


Figure 2.12: Relative risk functions when β=1 and ϕ =3 and resulting histograms when choosing RRF4.

	1982-1987	1988-2001
Number of wards	130	130
Number of cases	305	1050
Total person years	3331224	7772856
Average disease frequency*	9.15779	13.5085
Relative Risk Function	RRF4	RRF4
Test statistic	63.450	213.543
Monte Carlo p-value	0.370	0.550

** per 100,000 population*

Table 2.29: Bithell's Linear Risk Score analysis.

For hypothesis testing, from the p-values obtained, there was no evidence of clustering within NYG landfill site.

2.6.4. Diggle's Method

Diggle's method (1990) was redeveloped by Diggle and Rowlingson (1994). It uses case-control analysis to determine the level of possible clustering in the dataset. Individual level data are required for this method to represent the cases and controls as opposed to case and population at risk data as in the previous focused methods. The spatial pattern of the cases is compared with the controls. The spatial pattern of the controls acts as a null model of no clustering and should represent the population at risk.

Test statistic

The test statistic is based on maximising the likelihood of the cases and controls based on an exponential decline in risk as the squared distance from the focus increases. The relative risk function (RRF) used is shown in equation 2.13. The null hypothesis assumes no change in density of cases with respect to the focus; hence the RRF is equal to 1. The underlying model is that of an inhomogeneous Poisson process. The alternative hypothesis is that there is a raised density of points near the focus. The default value is β=0 (null hypothesis of no clustering) where β is the intercept and ϕ is the distance decay of the function. ρ represents the number of events per unit area and λ_{0_x} denotes the spatial variation of the controls irrespective of the focus.

$$\lambda_x = \rho\lambda_{0_x} RRF$$
$$RRF = 1 + \beta e^{-\phi d^2}$$

Equation 2.13: The relative risk function using Diggle's method.

Parameters are optimised via maximum likelihood estimation and the fit of the case data to the model is compared with a generalised likelihood ratio test (two models are used, one that has no relationship to the focus, the other that does have a relationship with the focus). ClusterSeer V2.2.4 finds the best fit of the initial values and displays both sets of parameters. The maximised likelihood (from the fitted model) and original likelihood

(from the initial values entered) are reported in the output. The model is evaluated with the generalised likelihood ratio, which compares the fitted model with the null hypothesis.

The generalised log likelihood ratio test evaluates which model better explains the data. The generalised log likelihood test statistic that Diggle and Rowlingson (1994) use to compare the models is shown in equation 2.14 where ρ is the overall number of cases per unit area, $L_a(\rho)$ is the log likelihood for the alternative hypothesis Ha and $L_0(\rho)$ is the log likelihood for the null hypothesis Ho. $P(c_{1i})$ is the probability that location i is the location of a case. The significance of the GLT is obtained using the chi-squared distribution with 2 degrees of freedom. The parameters are optimised via maximum likelihood estimation. ρ is maximised when the RRF is equal to 1 for a particular number of cases c_1 and particular number of controls c_2.

$$GLT = 2[L_a(\rho) - L_0(\rho)]$$
$$L_0(\rho) = c_1 \log \rho - (c_1 + c_2)\log(1+\rho)$$
$$L_a(\rho) = \sum_{i=1}^{n} P(c_{1i}) + \sum_{i=c_1+1}^{c_1+c_2} \log(1 - P(c_{1i}))$$
$$P(c_1) = \frac{\rho RRF}{1+\rho RRF}$$

Equation 2.14: The generalised log likelihood ratio test and log likelihoods

Results

Ho: The disease does not cluster around the point source.
Ha: The disease does cluster around the point source.

Using the leukaemia datasets within 20km of NYG landfill site and the same time periods as the previous focused methods, results are shown in table 2.30. The initial parameters

used were $\beta=1$ and $\phi=0.5$ (assuming that the relative risk is 2 at the focus and then decreases to 1 at a few kilometres away from the focus).

	1982-1987		1988-2001	
	p-value	β, ϕ, ρ	p-value	β, ϕ, ρ
Lower Body Cancers	0.9878	0.1404	0.0347	-0.3330
		0.6633		0.0403
		0.2425		0.2711
NHSAR Controls	0.7702	-0.9506	0.0573	-0.3251
		1.3176		0.0351
		0.5710		0.5434

2.30: p-values and fitted values obtained using Diggle's method.

Table 2.30 shows non-significant results for the earlier period 1982-1987 when using both control datasets with the cases but a significant result is found for the later period when using lower body cancers and a borderline significant result is obtained when using the NHSAR control dataset. The parameters β, ϕ and ρ of the fitted model are also shown in table 2.30. Choosing other values of β and ϕ give very similar p-values to those quoted in table 2.30 along with very similar resultant parameters for β, ϕ and ρ. Note that for the period 1988-2001, both values of β were less than zero indicating that, in fact, the relative risk is lower at NYG compared with the surrounding area.

Note that these results are different to the previous methods. This method analyses data at individual level, unlike the other methods which analyse the data at aggregated level (ward). This could be the reason why the results differ to previous methods.

It appeared that irrespective of the choice of β and ϕ entered into the model, the fitted model converges to very similar optimum values of β and ϕ. However, this is due to these datasets being very large. Another point source was analysed which was situated in a rural part of Wales. Varying the values of β and ϕ produced large differences in p-values and resultant β and ϕ. Thus, caution is advised regarding the choice of parameters if the dataset is small.

Table 2.31 summarises the p-values obtained for all focused methods using the leukaemia dataset in Wales 1982-2001.

	1982-1987	1988-2001
Kulldorff (population)	0.157*	0.673**
Kulldorff (Lower body cancers)	0.444*	no cluster
Kulldorff (NHSAR)	no cluster	no cluster
Score Test	0.201	0.667
Bithell's Linear Risk Score (Hypothesis testing)	0.370	0.550
Diggle (Lower body cancers)	0.988	0.035
Diggle (NHSAR)	0.770	0.057

* radius 1.73km ** radius 1.53km

Table 2.31: Summary of p-values obtained for focused methods.

The case and control data produced a significant and borderline significant result. All other results were non-significant. The significant result could be due to point data being analysed and not areal data as was the case for the other methods excluding Diggle's method. In general, the analysis for the case and population at risk data consistently produced non-significant results for all methods. However, the disadvantage of both the Score test and Bithell's test is that no actual "cluster" is detected. Observed and expected cases are plotted with an overall p-value but no "cluster" is found because the relative risk is being modelled with a given focus. They do not look for a cluster in the usual way; they assume the risk varies continuously. Additionally, the Score test and Bithell's method calculate a background rate based on the dataset used. i.e. the datasets used in this section were within 20km of the focus. If the dataset used was within 10km of the focus, for example, the background rate would change and a completely different p-value could be obtained.

2.7. Space-Time Clustering Methods

The previous section analysed the leukaemia dataset and detected clusters, depending on the true underlying situation and if the algorithms were sufficiently powerful. The dataset examined covered a period of twenty years and a cluster may exist for just a small number of years within the twenty year period. Space-time clustering methods may

identify local clusters but may also identify the specific time period in the dataset when the cluster existed.

2.7.1. Space-Time Scan Statistic (Kulldorff)

Kulldorff's spatial method was defined in section 2.5.3. The space-time scan statistic is an extension of this method. It uses a cylindrical window rather than a circular window for local and focused methods described earlier. The height of the cylindrical window defines the temporal element of the potential clusters. The cylinder is used as a generic volume analogous to the circles of the purely spatial method. The window is moved in space and time so that for each location and size it also looks at each possible time period (in years regarding the leukaemia dataset). The height of the cylinder is varied (the number of years in the cluster) along with the window over the spatial geographical units to find the most likely cluster. The full dataset of leukaemia cases for the period 1982-2001 was used for this method. A maximum cluster size of 5% of the total population at risk was used and a maximum temporal cluster of 50% was used (a maximum of a ten year cluster period) and was also the maximum value that could be entered.

Results

Ho: There is no space-time clustering in the dataset.
Ha: There is space-time clustering in the dataset.

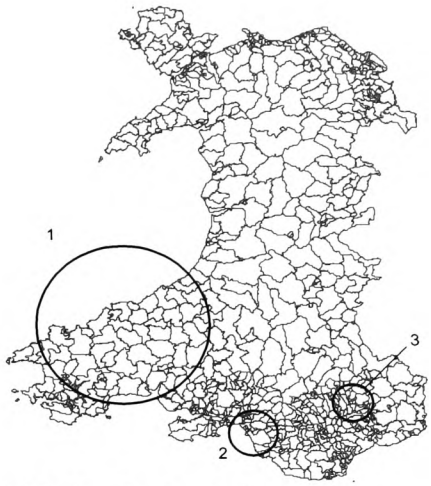
Table 2.32 locates the clusters using this method as a local method (identifying clusters in the entire study region i.e. Wales) and focused method (within 20km of NYG landfill site).

	Local Method			Focused Method		
	Population	NHSAR	LBC	Population	NHSAR	LBC
Cluster time period	1994-1999	2000-2001	1986-1995	1996	1996	1993
Number of wards in cluster	71	31	37	30	35	5
Coordinates of cluster	220650, 241300	277907, 188453	322088, 203201	NYG	NYG	NYG
Radius (km) of cluster	38.47	10.89	8.93	8.84	8.87	1.86
Population/person years	136749	94	874	139369	72	10
Observed cases in cluster	196	61	240	34	35	5
Annual cases/100000 population	23.9	NA	NA	24.3	NA	NA
Expected Cases in cluster	110.11	31.33	168.86	18.79	24.00	1.93
Observed/Expected in cluster	1.780	1.947	1.421	1.810	1.458	2.588
p-value of cluster	0.001	0.001	0.004	0.217	0.589	0.926

Table 2.32: Kulldorff's space-time analysis as a local and focused method.

Table 2.32 shows significant results when the local method is used to detect clusters in Wales. All three methods – population at risk, lower body cancer controls (labelled as LBC in table 2.32) and NHSAR controls locate significant clusters but in different areas of Wales along with different time periods. A large cluster was located in West Wales using the case and population at risk data as a local method, probably too big to term a cluster due to its geographical size. The case and population at risk data along with the NHSAR controls both produced very similar sized clusters within NYG for the single year 1996 but were not significant. However, the lower body cancer controls identified the year 1993 as the most likely cluster but was highly non-significant at p=0.926, more evidence suggesting that lower body cancers may not have been appropriate as controls for the leukaemia cases. Further work is warranted regarding the results shown above. It may be worthwhile extracting the corresponding observed cases and expected cases using the case and population data in the context of the cluster areas found using the case and control methods to identify any possible reasons as to why results are different.

Figure 2.13 locates the clusters for the local space-time clustering algorithm.



1 Population at risk data 2 Lower body cancers 3 NHSAR controls

Figure 2.13: Location of clusters using the space-time local clustering algorithm.

Table 2.33 shows a summary of the space-time cluster algorithms along with p-values obtained using the leukaemia dataset in Wales for the period 1982-2001. Kulldorff's local method produced significant results in West Wales, Swansea and Torfaen. Kulldorff's focused statistic failed to locate a significant cluster around NYG landfill site.

	p-value	period	comments
Kulldorff (local,population)	0.001	1994-1999	radius 38.470km in West Wales
Kulldorff (local, LBC)	0.001	2000-2001	radius 10.890km in Swansea
Kulldorff (local, NHSAR)	0.004	1986-1995	radius 8.390km in Torfaen
Kulldorff (focused, population	0.217	1996	radius 8.84km from NYG
Kulldorff (focused, LBC)	0.589	1996	radius 8.87km from NYG
Kulldorff (focused, NHSAR)	0.926	1993	radius 1.86km from NYG

LBC – lower body cancers

Table 2.33: Summary of space-time clustering algorithms.

All other space-time methods use case data only and these methods were disregarded from the outset due to the fact that the population varies markedly between wards in Wales.

In summary, table 2.34 provides a summary of the advantages and disadvantages of the clustering algorithms examined.

GLOBAL METHODS		
Cluster Algorithm	**Advantages**	**Disadvantages**
Moran's I statistic	Detects departures from spatial randomness.	Population at risk not accounted for – thus large differences in populations between geographical units will decrease the ability to detect true clustering.
Oden's I Pop	Detects departures from spatial randomness by taking population into account.	No locations are identified; this is a global method only.
Besag and Newell	Gives a general "picture" of where clustering occurs. Gives an overall "risk" of observed to expected number of clusters.	Sensitive to changes in the maximum cluster size. The significance of a local cluster depends on the level of aggregation and the chosen value of the maximum cluster size k (Waller and Turnbull, 1994)
Cuzick and Edwards'	Identifies whether there is clustering in the dataset or not using case and control data. Can take confounding into account.	Does not indicate any locations of clustering. The algorithm takes longer to run the higher the value of the maximum number of nearest neighbours k.

LOCAL METHODS		
Cluster Algorithm	**Advantages**	**Disadvantages**
Besag and Newell	Gives a general "picture" of where clustering occurs. Gives an overall "risk" of observed to expected number of clusters.	Sensitive to changes in the maximum cluster size. The significance of a local cluster depends on the level of aggregation and the chosen value of k (Waller and Turnbull, 1994)
Kulldorff's Spatial Scan Statistic	Not as sensitive to changes in the maximum population size as other algorithms. Allows local spatial clustering, focused clustering, local space-time clustering and focused space-time clustering. Can take confounding into account.	Tends to aggregate nearby clusters into one large cluster.
Anselin's Local Moran Test	Can be used as a diagnostic for outliers in the dataset. Can be used as an indicator of local spatial clusters.	Does not take population at risk into account. Areas with large population differences will bias results.
Turnbull's Method	Identifies the three most likely clusters. Should be used when population at risk is known a priori.	The three most likely clusters may overlap. Tends to aggregate nearby clusters into one large cluster.

FOCUSED METHODS		
Cluster Algorithm	**Advantages**	**Disadvantages**
Score Test (Lawson and Waller)	Each region is weighted by degree of exposure to the focus.	Tends to aggregate nearby clusters into one large cluster. Resultant p-values depend on the distance that the user decides to test the hypothesis.
Bithell's Linear Risk Score Test	Sensitive to excess risk near a point source. Four models are available to model the data.	RRF parameter should be chosen objectively. P-values cannot be interpreted if parameters are chosen to fit the relevant model rather than hypothesis testing. Resultant p-values depend on the distance that the user decides to test the hypothesis.
Kulldorff's Spatial Scan Statistic	Not as sensitive to changes in the maximum population size as other algorithms. Allows local spatial clustering, focused clustering, local space-time clustering and focused space-time clustering. Can take confounding into account.	Tends to aggregate nearby clusters into one large cluster.
Diggle's Method	Can take confounding into account.	Significance of tests can vary depending on the initial values of ϕ and β that are input by the user if dataset is small.

SPACE-TIME METHODS		
Cluster Algorithm	**Advantages**	**Disadvantages**
Kulldorff's Space-Time Scan Statistic	Uses case and population data or case and control data. Allows local spatial clustering, focused clustering, local space-time clustering and focused space-time clustering. Can take confounding into account.	Tends to aggregate nearby clusters into one large cluster.

Table 2.34: Summary of clustering algorithms.

2.8. Literature Review

The literature review is presented here since it was deemed necessary that the clustering algorithms be defined in detail so that the algorithms can be understood.

Background

Cancer is generally caused by a combination of factors that interact in ways which are not as yet fully understood. Some cancers may not be caused by one factor but by a combination of many factors over time. Environmental factors such as air and water quality account for less than 10% of most cancers (Perera, 1997). Diet has been estimated to account for 80% of cancers of the large bowel, breast and prostate (Doll et al., 1981; Willett, 1995). It is estimated that around half of all cancers are behavioural and are thus potentially controllable (Vale of Aylesbury Primary Care Trust NHS report 2005/06[4]). Some people are predisposed to certain types of cancer due to genetics. Cancers of the thyroid and leukaemia have been typically linked with radiation (Jaworowski, 1999). There has been a vast amount of research conducted to examine the possibility of a raised risk of childhood cancer in the vicinity of certain nuclear

[4] http://www.bucksss.nhs.uk/NHSAnnualReport0506.pdf

installations which could in turn be associated with radioactivity released from these sites (e.g. Black, 1984; Gardner et al., 1990). There has been a significant amount of research focused on identifying clusters of leukaemia. (e.g. Hjalmars et al (1996), Cartwright et al (1990), Murrin et al (2005)).

Occupational exposure to radiation and benzene has been identified as a potential risk factor for leukaemia (Yin et al., 1996). Studies have also suggested electromagnetic fields (EMFs) and residence to electric power lines as potential risk factors. Ionising radiation is known to be a carcinogen. In 1997, a UK study by Knox and Gilman linked childhood leukaemia to oil refineries and railway yards.

However, here, we are not reviewing the algorithms in terms of application to cancer but aim to review specific algorithms that have been examined in the previous section and detailing past studies that have compared more than one algorithm.

Studies investigating clustering methods tend to investigate a cancer of high incidence in order for the study to have a large number of registrations and enabling high power in detecting clusters, if any clusters exist.

Studies analysing cancer data using one clustering algorithm

The Kulldorff Scan Statistic was used in a study by Hjalmers *et al.* (1996) who analysed 1523 cases of acute childhood leukaemia (0-15 years) in Sweden for the period 1973-1993. No significant results were obtained when analysing acute lymphoblastic leukaemia (ALL) or acute non-lymphoblastic leukaemia (ANLL) or both together. The analysis period covered 21 years of data – it was possible that a cluster existed for earlier or later periods in that particular dataset - a space-time scan statistic could have analysed this data further.

Kulldorff's Spatial Scan Statistic (Kulldorff 1997, 2004) was used as a surveillance tool by Kulldorff (2001) to survey a current cluster of disease and adjust for multiple testing

due to various locations and sizes of disease cluster tests. The dataset consisted of all male thyroid cancers diagnosed in New Mexico for the time period 1973-1992. The spatial scan statistic initially analysed the six-year period 1973-1978. The spatial scan statistic was repeatedly used with an additional year of data until the whole period 1973-1992 was analysed. However, since the whole time period was analysed and a true risk may have been present during the last few years a space-time scan statistic should have been used. Again, 1973-1978 was initially analysed followed by an extra year of data to identify whether the same cluster was identified when including another year of data. Most results were non significant; however when the diagnosis period 1973-1991 was analysed, Los Alamos was identified as a significant cluster, p=0.02 for the specific space-time period 1989-1991. With the following year added to the analysis, the same area was identified for the period 1989-1992, p=0.002. However, the p-values were not adjusted for the multiple time period analysis performed over all the years. When the p-values were adjusted the result for the period 1989-1991 was non significant, p=0.13 but the period 1989-1992 was still statistically significant, p=0.013. The latency period is important with this type of analysis since it was known that exposure to known risk factors had occurred many years before diagnosis. Thus exposed people may have moved away and unexposed people may have moved in, known as population mixing.

Anselin's Local Moran test was implemented by Jacquez and Greiling (2003) who examined local clustering in breast, lung and colorectal cancer in Long Island, New York for the five year period 1993-1997. The data were aggregated to zip code at diagnosis. Anselin's local Moran test was used to identify significant clustering and spatial outliers using the ClusterSeer V2.2.4 software. An adjusted level of significance was used to allow for multiple testing due to shared neighbours and the fact that local clusters could overlap. Results showed various clusters of standardised morbidity ratios (SMR) for the three cancers studied; these were compared with the overall New York rates. The overall figures and rates were not discussed – thus no clear comparison could be made. Some clusters contained few cases; thus a decrease of just one or two cases could dramatically affect the significance of that particular cluster. As always with cancer studies, cancer latency, the time between exposure and onset of cancer are important factors. For the

cancers studied, cancer latency varies between five and forty years. Thus the zip code at diagnosis may not necessarily have been the location where the cancer developed. Additionally the geography of Long Island is unlike other areas since it is a long forked island – ultimately it is poor to analyse this island using circles and centroids. In recent years, Kulldorff has created an elliptical scan statistic which may have been useful here.

Studies analysing cancer data using two or more clustering algorithms

Bellec *et* al (2006) used Moran's method (to determine spatial autocorrelation, if any), Knox's method (this method identifies the number of pairs of cases that are separated by less than the critical time and critical space values that are input by the user) and Kulldorff's Space and Space-Time Scan Statistic (to scan the whole area for local clusters) using a dataset of childhood acute leukaemia in France for the period 1990-2000. A statistically significant spatial heterogeneity of a very small magnitude was observed in the incidence of childhood leukaemia for the period 1990–1994 only (no significant results for the whole time period or the second period 1995-2000). Cases older than 10 years living in the same area at diagnosis tended to cluster within 6 months (using the Knox test). Caution is advised when using Knox's method to determine the choice of critical values since these can affect resulting p-values.

Various clustering algorithms have been rigorously reviewed by Anselin (2004) in conjunction with the GIS Ad Hoc Committee for North American Association of Central Cancer Registries (NAACCR). The review was based on factors such as free software, latest versions, development, documentation, downloadability and the operating system required. The NAACCR consider the report to be a valuable resource in the difficult area of clustering studies. Anselin states that an effective clustering algorithm should include efficient data input and flexible output including visualisation of results. The program must offer descriptive spatial statistics as well as point pattern analysis and spatial autocorrelation analysis. Two methods Anselin states that were essential in locating clusters were the Local Moran test (biased since it is his own method) and Kulldorff's spatial scan statistic. Clustering algorithms ultimately differ slightly in their function,

methodology and usage. It is worth noting the difference between analysis by data points and analysis by areal units. For point data, the algorithms determine the extent to which other points (i.e. cases and not controls) are closer than they would be in a reference situation. For areal data, algorithms determine whether an areal unit has similar units surrounding it than would occur randomly (spatial autocorrelation). This report stated that Kulldorff's spatial scan statistic was currently the most widely used algorithm in the public health departments to detect disease clusters in North America.

Simulated studies of clustering

Kulldorff *et al* (2005) analysed three clustering algorithms, the Kulldorff Scan Statistic, the maximised excess events test (MEET) (Tango, 1995) and the non-parametric M statistic (Bonetti-Pagano, 2001, 2005). MEET is the maximised test of a weighted sum of excess events based on two distance based exponential weight functions. The non-parametric M statistic assumes a Gaussian process and measures the distance between a function based on the observed case locations and a function based on the population at risk. These functions were based on distances between pairs of individuals. Nearly one and a quarter million benchmark datasets were generated under fifty-one cluster models. The female population in 245 counties in North-eastern United States of America was used to generate the benchmark datasets (using the 1990 Census). Two groups of 100,000 datasets were analysed with 600 and 6000 simulated cases each. It was shown that all three tests had high power, depending on the area under investigation, these areas being urban, rural or mixed (a large city surrounded by rural areas). The Kulldorff Scan Statistic had highest power in detecting a relative risk of 27.03 in a rural area hot spot cluster at 0.991 for a cluster of 1 county, 6000 simulated cases and significance level 0.05 (the power was 0.936 for a cluster containing 16 counties with a relative risk of 3.90). MEET had highest power in an urban area for a hot spot cluster of 1 county, 6000 simulated cases and significance level 0.05 (the power was 0.982 for a cluster of 16 counties). The non-parametric M statistic was shown not to have the highest power for any type of area in the study. The Kulldorff Scan Statistic was best used in detecting localised clusters while the MEET was better at detecting global clustering present

throughout the study region. The Kulldorff Scan Statistic can only detect 'circular-shaped' clusters; this method would be unable to detect a cluster along a power line, if one existed.

The same benchmark datasets as the previous study were used by Song *et al.* (2003) in eight clustering tests – although only 100,000 datasets were used with 600 simulated cases. The power of the spatial scan statistic (suited to detecting hot spot clusters) and the MEET (suited to detecting global clustering) were compared with six other tests: Besag and Newell, Cuzick and Edwards', Swartz entropy (Swartz, 1998), Whittemore's test (Whittemore et al., 1987), Moran's I and modified Moran's I (Song et al., 2003). Swartz' cluster detection test used the method of entropy as the test statistic. Whittemore's test was based on the product sums of the distances between all two counties i and j and the number of cases in counties i and j. Results show that the spatial scan statistic and the MEET both performed well most of the time. However with the right choice of parameters, Besag and Newell's method had high power for a mixed area, 0.983 for a cluster of 1 county whereby the relative risk in the simulated cluster was 2.85, significance 0.05, and 0.956 for a cluster of 8 counties whereby the relative risk in the simulated cluster was 2.24 (compared with 0.936 and 0.941 respectively with the spatial scan statistic) for 8 counties.

2.9. Applying simulated datasets to the algorithms

"Testing space-time and more complex hyperspace geographical analysis tools" by Openshaw *et al* (2000) compares the performance of several exploratory geographical analysis methods using a number of simulated datasets that contain various spatial patterns as described in the literature review. The user is invited to use their datasets to test and develop new methods of geographical analysis.

Various datasets have been selected from Openshaw *et al* to test the clustering algorithms in ClusterSeer V2.2.4 and SaTScan V5.1.3 and to investigate the accuracy of the analysis of these methods. Four datasets were obtained from Openshaw's datasets and four datasets were obtained from Brunsdon's datasets to analyse. The datasets are available to

download at the Centre for Computational Geography at Leeds[5]. The study is also available at this website.

The simulated datasets are of case (event) data and corresponding population at risk. Those algorithms that will be compared to each other are Turnbull v Kulldorff and Score Test of Lawson and Waller v Kulldorff (those algorithms that performed the best from the earlier analysis and that use case and population data).

2.9.1. Openshaw's datasets

The datasets correspond to dataset 02, dataset 03, dataset 04 and dataset 05 of Openshaw's files.

All datasets consist of population data from the Yorkshire and Humberside region from 10,430 Census Enumeration Districts (ED). 1000 events were generated from a total population at risk of 4,820,129 persons (20.75 per 100,000 population). The datasets by Openshaw had varying degrees of clustering and different parent locations (clusters being part of the 1000 events). Multinomial allocation and multinomial probabilities proportional to the population at risk were used to select the random cases. An inhomogeneous Poisson process reflecting a Gaussian risk function was used by Openshaw et al. (2000) to determine the cases in the clusters. Figure 2.14 details the location of clusters in dataset 04 (rates per 100,000 population). Table 2.35 shows the number of cases and radii of the clusters in the datasets used. Note in dataset 04 that clusters of size 1 are shown. This is due to a very high rate in the population at risk. In reality, a cluster of size 1 would not be deemed a cluster as one case in an area cannot be classed as a cluster. It should be noted that the rates found in these clusters are very high compared to the background rate in the simulated dataset.

[5] http://www.ccg.leeds.ac.uk/software/smart/data/

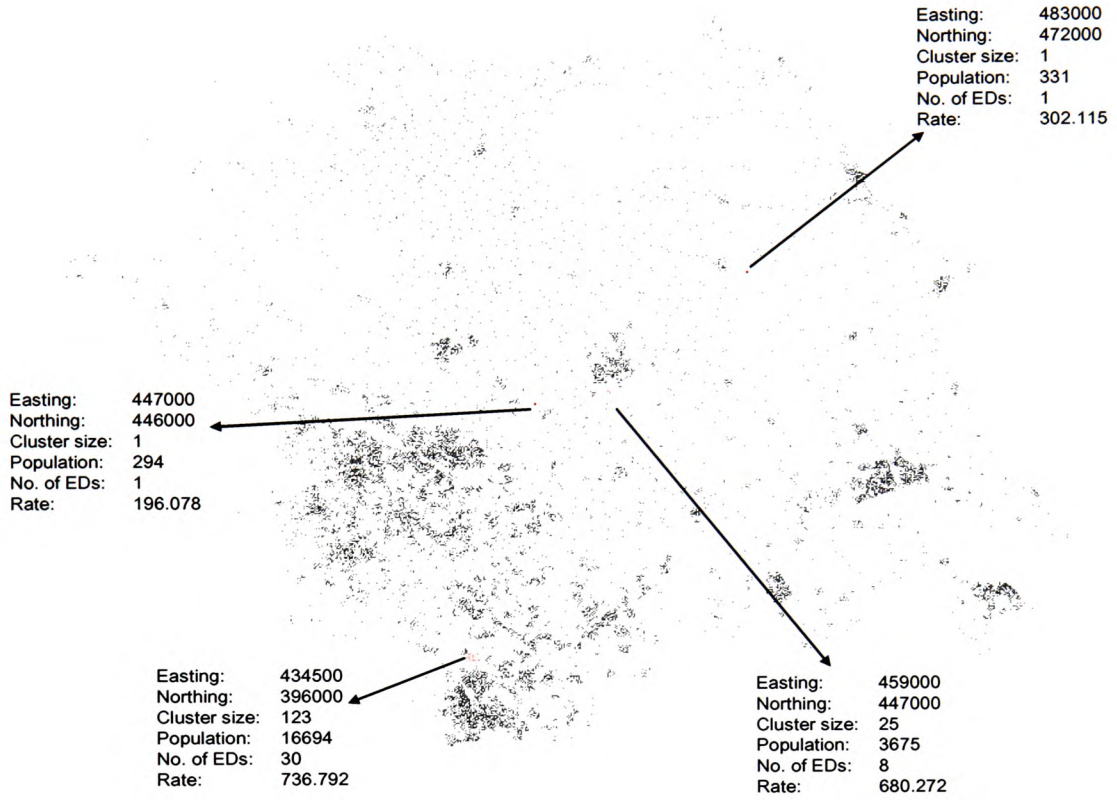


Figure 2.14: Location of clusters for dataset 04.

	Coordinates (E, N)*	Number in cluster	Radius	Rate**
Dataset 02	434500, 396000	200	1.74km	1065.587
Dataset 03	No cluster	No cluster		No cluster
Dataset 04	434500, 396000	123	1.41km	736.792
	483000, 472000	1	1.84km	302.115
	447000, 446000	1	1.25km	196.078
	459000, 447000	25	1.70km	680.272
Dataset 05	434500, 396000	87	1.45km	486.986
	459000, 447000	13	0.71km	353.741
Background rate = 20.75 per 100,000 population				

* *E-Easting, N-Northing,*

** Rate *per 100,000 population*

Table 2.35: Centroids and cluster sizes for datasets used.

Appendix C identifies the cumulative relative risk for each of the clusters in the datasets.

2.9.1.1. Turnbull's method v Kulldorff's spatial scan statistic

Various population sizes were entered to locate possible clusters in all datasets using Turnbull's method. There were 1000 events in the dataset covering a population of 4820129, hence 1 event for approximately every 4820 population. Table 2.36 summarises the analysis for Turnbull's method. The table shows the p-values obtained and observed numbers in brackets for the three most likely clusters for various maximum population sizes (noting that sometimes the observed number may not be a whole number since this method applies a proportion to population at risk and cases in a ward to obtain the exact maximum population size).

	DATASET 02			DATASET 03		
Maximum population	**Cluster 1**	**Cluster 2**	**Cluster 3**	**Cluster 1**	**Cluster 2**	**Cluster 3**
4800	0.001(86.9)	0.001(86.3)	0.001(84.7)	0.878(6.0)	1(5.0)	1(4.9)
9600	0.001(143.2)	0.001(142.4)	0.001(141.9)	0.954(8.0)	0.956(8.0)	0.976(7.3)
24000	0.001(204.0)	0.001(203.5)	0.001(203.0)	0.957(13.0)	0.971(12.3)	0.972(12.3)
48000	0.001(207.0)	0.001(206.0)	0.001(205.8)	0.857(20.0)	0.974(20.0)	0.975(19.9)
144000	0.001(227.0)	0.001(226.0)	0.001(225.0)	0.904(45.0)	0.907(44.8)	0.968(44.0)
240000	0.001(244.0)	0.001(243.0)	0.001(242.7)	0.588(71.0)	0.589(70.9)	0.589(70.9)
600000	0.001(310.0)	0.001(309.9)	0.001(309.0)	0.917(147.0)	0.949(146.0)	0.97(145.0)
	DATASET 04			**DATASET 05**		
Maximum population	**Cluster 1**	**Cluster 2**	**Cluster 3**	**Cluster 1**	**Cluster 2**	**Cluster 3**
4800	0.001(45.1)	0.001(44.6)	0.001(43.0)	0.001(33.6)	0.001(33.1)	0.001(32.6)
9600	0.001(83.5)	0.001(83.1)	0.001(82.6)	0.001(62.9)	0.001(62.5)	0.001(61.7)
24000	0.001(128.0)	0.001(126.9)	0.001(126.1)	0.001(92.0)	0.001(91.0)	0.001(90.2)
48000	0.001(133.0)	0.001(132.0)	0.001(131.6)	0.001(97.0)	0.001(96.0)	0.001(95.6)
144000	0.001(153.0)	0.001(153.7)	0.001(152.0)	0.001(117.0)	0.001(116.0)	0.001(116.0)
240000	0.001(175.0)	0.001(174.0)	0.001(173.0)	0.001(140.0)	0.001(139.0)	0.001(138.0)
600000	0.001(235.0)	0.001(234.0)	0.001(233.4)	0.001(205.0)	0.001(204.0)	0.001(203.4)

Table 2.36: P-values (observed figures in brackets) obtained using Turnbull's method.

Dataset 02

All clusters located using Turnbull's method and dataset 02 were in the correct area (i.e. the resulting cluster area was in the area where the actual cluster was located) but had varying observed figures in the clusters due to the maximum population level used.

Dataset 03

No significant clusters were found when analysing this dataset. This is correct since no actual clusters were generated in this dataset. Excluding a population size of 240,000 persons, all p-values were greater than 0.857 for the most likely cluster. For a population size of 240,000 persons, three most likely clusters were found. None of the clusters were significant but were located in the lower right hand corner of the study region even though there was no evidence for a cluster being situated there. All clusters overlapped – a major disadvantage with this method. For a population size of 48,000 persons, the most likely cluster contained a rate of 29.5833 per 100,000 persons whereas the whole area contained a rate of 20.7463 per 100,000 persons, an approximate 43% increase compared with the entire study region, however this result was non-significant at p=0.857.

Dataset 04

Using dataset 04, all clusters found were significant for all population sizes. The locations of the clusters were similar for all maximum population sizes in terms of the centroid of the cluster - the only difference being the radius of the cluster due to the size of the population used. All clusters contained the area of the cluster of size 123. The second and third most likely clusters also contained the same area – geographic overlap was evident. Three of the results located a cluster in an area other than that where the actual clusters were found. However, by looking at table 2.36, the clusters of size 1 would have to have had very small maximum population sizes in order for them to have been found (i.e. they would not have been found with the population sizes used). To summarise, Turnbull's method identified the largest cluster for all maximum population sizes. However, all clusters were located in the same area due to overlapping and no other clusters were located.

Dataset 05

The most likely cluster was located in the correct area for two of the population sizes. One of the population sizes located a cluster in a completely different area due to the overlapping clusters that Turnbull produced.

To compare Turnbull's algorithm with Kulldorff's method the maximum population size was compared directly to Turnbull's method. i.e. a population size of 240,000 persons corresponds to 4.979% of the total population at risk. Table 2.37 summarises Kulldorff's spatial scan statistic for various population sizes. Note that Kulldorff's method does not produce clusters that geographically overlap, hence no observed cases in one cluster are obtained in another cluster, unlike Turnbull's method. However, Kulldorff's method tends to have the correct number of observed cases in a cluster irrespective of the choice of the maximum population size.

	DATASET 02			DATASET 03		
Maximum population	**Cluster 1**	**Cluster 2**	**Cluster 3**	**Cluster 1**	**Cluster 2**	**Cluster 3**
0.100%	0.001(86)	0.001(37)	0.976(24)	0.983(2)	0.995(5)	0.997(2)
0.199%	0.001(140)	0.997(29)	0.997(9)	0.99(2)	0.996(5)	0.999(2)
0.498%	0.001(199)	0.997(4)	0.997(5)	0.996(2)	-	-
0.996%	0.001(199)	0.997(4)	0.997(5)	0.996(2)	-	-
2.988%	0.001(199)	0.997(4)	0.997(5)	0.996(2)	-	-
4.979%	0.001(199)	0.997(4)	0.997(5)	0.997(2)	-	-
12.448%	0.001(199)	0.997(4)	0.997(5)	0.997(2)	-	-
	DATASET 04			**DATASET 05**		
Maximum population	**Cluster 1**	**Cluster 2**	**Cluster 3**	**Cluster 1**	**Cluster 2**	**Cluster 3**
0.100%	0.001(41)	0.001(40)	0.001(24)	0.001(32)	0.001(13)	0.001(15)
0.199%	0.001(83)	0.001(24)	0.001(20)	0.001(62)	0.001(13)	0.001(18)
0.498%	0.001(123)	0.001(24)	0.994(4)	0.001(87)	0.001(13)	0.995(4)
0.996%	0.001(123)	0.001(24)	0.994(4)	0.001(87)	0.001(13)	0.995(4)
2.988%	0.001(123)	0.001(24)	0.994(4)	0.001(87)	0.001(13)	0.995(4)
4.979%	0.001(123)	0.001(24)	0.995(4)	0.001(87)	0.001(13)	0.995(4)
12.448%	0.001(123)	0.001(24)	0.995(4)	0.001(87)	0.001(13)	0.995(4)

Table 2.37: P-values (and observed cases in brackets) obtained using various population sizes using Kulldorff's spatial scan statistic for the three most likely clusters.

Dataset 02

Kulldorff's method correctly identified the most likely cluster in this dataset for population sizes over 0.199% of the entire population at risk. However, for small population sizes, the cluster found was smaller than the actual cluster. i.e. the maximum population at risk was too small to locate such a large cluster (in terms of population size). Two other secondary clusters were found but were non-significant.

Dataset 03

No significant clusters were found in this dataset which is consistent with Turnbull's method. No clusters were generated in this dataset; hence this method produced the correct results.

Dataset 04

Similar to dataset 02, when the maximum population sizes in the clusters were small, the analyses located smaller sized clusters but in the correct area where the actual clusters were found. For larger maximum population sizes, significant clusters of size 123 and 24 were found. The second cluster with 24 cases is slightly less than the actual sized cluster at 25. However, neither cluster of size 1 was found using this method due to the maximum population size being too large to detect such small clusters.

Dataset 05

As with other datasets, when the maximum population size was small, Kulldorff's method produced smaller sized clusters but in the area where the actual clusters should have been found. For larger maximum population sizes, this method correctly identified significant clusters of 87 and 13.

Figure 2.15 shows the actual locations of the clusters in dataset 02. The red circle identifies the actual cluster that is present in the dataset. The most likely clusters using Kulldorff's method are identified in blue for the various population sizes and the most likely clusters using Turnbull's method are identified in green for the various population sizes.

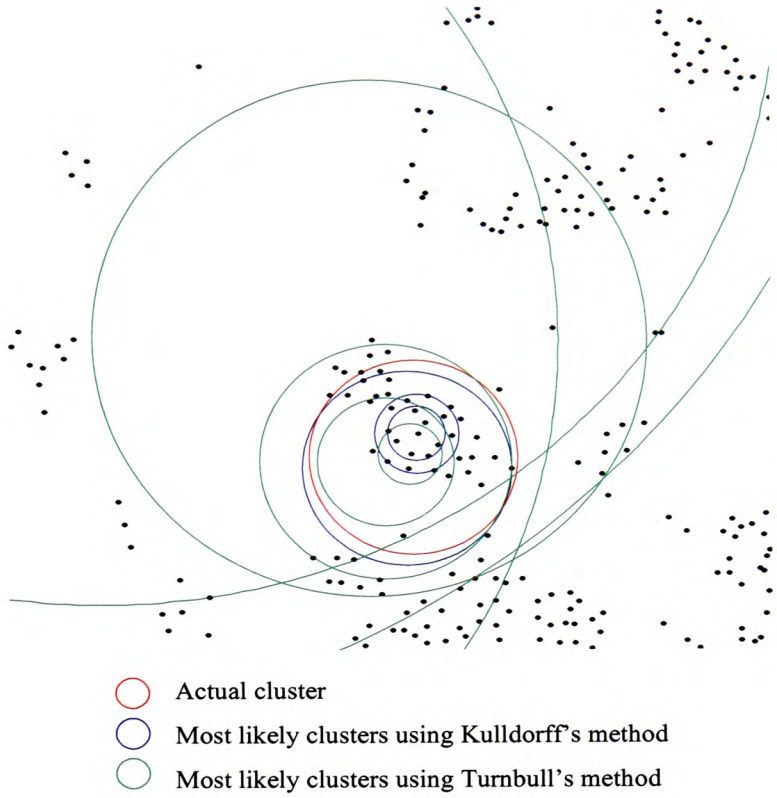


Figure 2.15: Locations of clusters using both methods for varying population sizes with dataset 02.

For Kulldorff's method, all population sizes over 0.498% produced the same sized cluster as the most likely cluster (the same was true for dataset 04). The main disadvantage using Turnbull's method that can be seen in figure 2.15 is the size of the most likely cluster when increasing the maximum population size. This is also true for the other three clustered datasets. Thus, to use Turnbull's method, the approximate population at risk must be known before analysis since the resulting number of cases (and size) in the cluster may be very large.

In summary it appears that Kulldorff's method appears to identify the correct clusters for the majority of maximum population sizes and no geographic overlap whereas Turnbull's method has the disadvantage in that the maximum population level must be reached before the numbers of cases are identified as being in a cluster. Turnbull's method

correctly identified the main cluster area of size 123 for all population values considered but of varying radii depending on the maximum population entered. However, this was the only cluster located as the three most likely clusters in the dataset due to overlap. Kulldorff's spatial scan statistic correctly located the majority of the actual clusters for most of the population sizes. Hence, this method was most accurate at correctly identifying clusters compared with Turnbull's method.

2.9.1.2. Kulldorff's spatial scan statistic v Score Test of Lawson and Waller

This section compares Kulldorff with the Score test as a focused test. Both methods are focused tests of clustering; coordinates are required for the centroid of a perceived cluster. For the Score test, datasets 02, 03, 04 and 05 were analysed within 20km of each of the foci as in the previous analysis. No clustering was found in dataset 03 using the Score test which is correct since no clusters were generated in the actual dataset.

Dataset 04 was used to analyse the two foci that consisted of clusters greater than size 1. The p-values obtained for both foci using the Score test were 0.001 indicating significant evidence of clustering at these foci. The histograms in figure 2.16 support the evidence of significant clustering (the test statistics are the extreme black lines). The two plots on the left show the observed and expected number of events within the foci. These plots clearly show the excess of observed cases to expected cases within approximately 1.5km of the foci and it can be seen that the observed and expected numbers stop diverging after this distance.

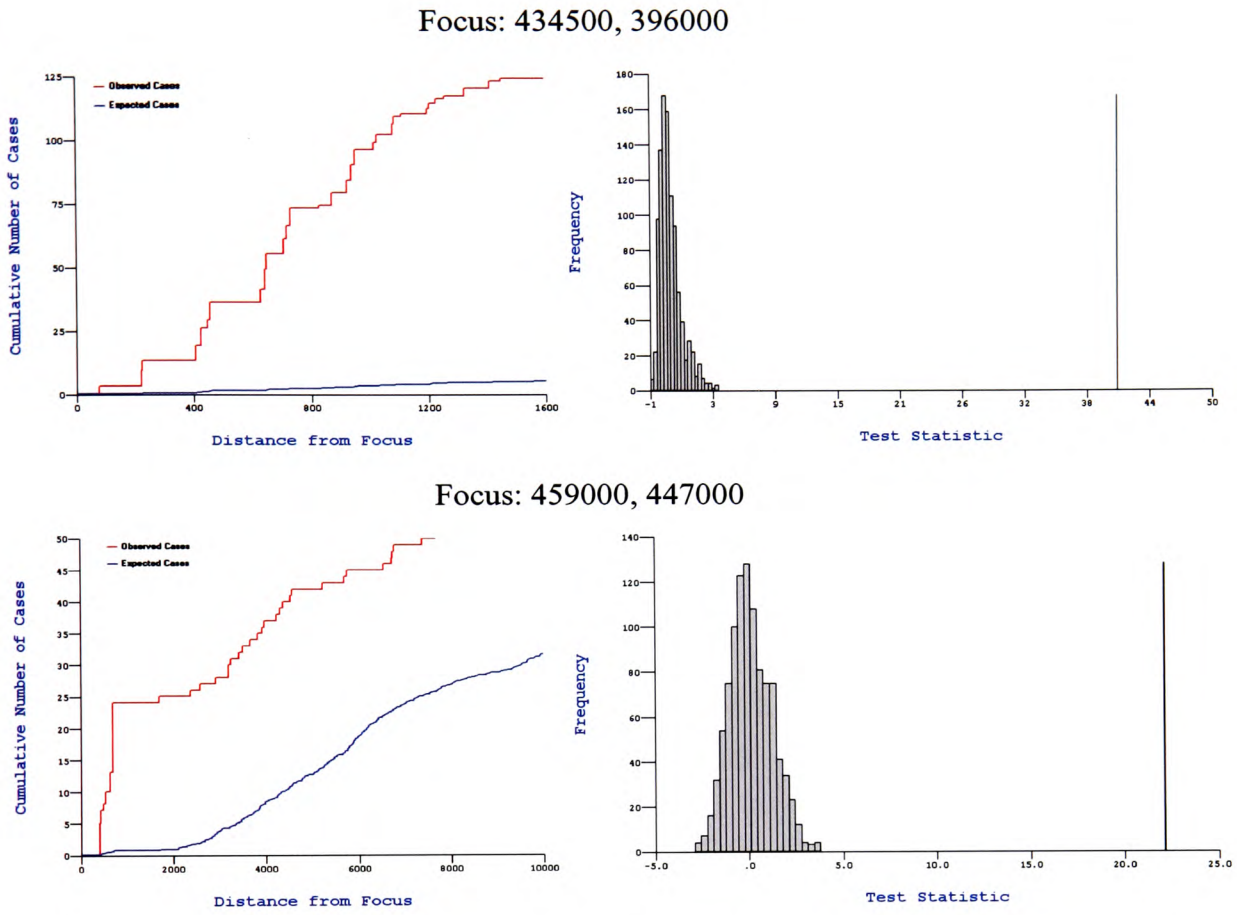


Figure 2.16: Plots and histograms for the foci using dataset 04 for the Score Test.

Dataset 02 and dataset 05 produced very similar results giving p-values of 0.001 at each of their respective foci and showing very similar excessive observed cases near the focus compared with the observed cases.

Table 2.38 shows the analysis of both datasets using Kulldorff's spatial scan statistic as a focused method using case and population at risk data.

Focus	Minimum Population Size	Dataset 02			Dataset 03		
		radius	p-value	cases	radius	p-value	cases
434500, 396000	0.50%	1.612km	0.001	199	1.912km	0.775	6
459000, 447000	0.50%	Only 1 focus in dataset 02			4.564km	0.791	13
434500, 396000	5%	1.612km	0.001	199	1.912km	0.533	6
459000, 447000	5%	Only 1 focus in dataset 02			No cluster identified		
434500, 396000	50%	1.612km	0.001	199	1.912km	0.898	6
459000, 447000	50%	Only 1 focus in dataset 02			44.554km	0.531	442

Focus	Minimum Population Size	Dataset 04			Dataset 05		
		radius	p-value	cases	radius	p-value	cases
434500, 396000	0.50%	1.414km	0.001	123	1.452km	0.001	87
459000, 447000	0.50%	0.707km	0.001	24	1.700km	0.001	14
434500, 396000	5%	1.414km	0.001	123	1.452km	0.001	87
459000, 447000	5%	0.707km	0.001	24	1.700km	0.001	14
434500, 396000	50%	1.414km	0.001	123	1.452km	0.001	87
459000, 447000	50%	0.707km	0.001	24	1.700km	0.001	14

Table 2.38: Focused test results using Kulldorff's spatial scan statistic.

Dataset 02

Kulldorff's focused test identified the correct area where the cluster of 200 cases was found but produced 199 cases in the cluster for each of the maximum population sizes entered.

Dataset 03

Clusters were located in the dataset for various population sizes but were not significant. This is correct since no clustering was generated in this dataset.

Dataset 04

Kulldorff's focused test identified the correct areas where the clusters of 123 cases and 25 cases were found but produced 24 cases in the cluster for each of the maximum population sizes entered for the actual cluster of 25 cases. However, Kulldorff's method correctly identified the cluster of 123 in all maximum population sizes.

Dataset 05

Significant clusters of size 87 and 14 were found in this dataset. These were in the area where the two actual clusters were located of sizes 87 and 13. Hence, one extra case was identified in one of the clusters compared to the actual clusters.

To summarise, Kulldorff's method located most of the actual clusters as a focused method. Dataset 03 produced non-significant clusters i.e. the rate found in these clusters were higher than the background rate but were not significantly higher. The disadvantage of the Score test is that only p-values are provided and a cumulative observed and expected chart. The number of observed cases in a cluster is not provided, only the cumulative number of cases observed from a specific focus although it should be able to be estimated using the observed and expected curves in the plot.

2.9.2. Brunsdon's datasets

The datasets refer to the same area as the previous datasets; population data from the Yorkshire and Humberside region from 10,430 Census EDs and 1000 events were generated from a total population at risk of 4,820,129 persons. These datasets used a more detailed model to select the cases for each of the datasets. The model used to generate the datasets is given in detail in Openshaw et al (2000). Essentially, the datasets were based upon three parameters; the number of parent locations, the percentage of clustered cases and a dispersion parameter (this differs according to the number of parent locations). A risk function was derived and locations of clustered cases were selected by multinomial allocation. All clusters were positioned at the same locations in all four datasets but the "intensity" of the clusters varied for each dataset. The overall rate in the study region for each dataset was 20.746 per 100,000 population. Dataset 2 and dataset 4 showed a lower rate than this for cluster 1. Thus, cluster 1 in dataset 2 and dataset 4 are not clusters of high rates but clusters of low rates. These were deliberately put in to identify whether the cluster detection was prone to false positive results. Also, it should be noted that the relative risks for some of these clusters were very high indeed. Therefore, similar to Openshaw's datasets the algorithms should have been able to locate these clusters without much of a problem.

Table 2.39 and figure 2.17 locate the clusters for the four datasets using Chris Brunsdon's simulated datasets. Appendix C shows the cumulative relative risk from the focus of the clustered datasets.

	DATASET 1			DATASET 2		
	Cluster 1	**Cluster 2**	**Cluster 3**	**Cluster 1**	**Cluster 2**	**Cluster 3**
Centroid	414500, 418750	461800, 495700	436950, 493540	414500, 418750	461800, 495700	436950, 493540
Radius	5km	5km	5km	20km	20km	20km
Cases	76	19	81	231	52	80
Population	148069	165	18255	1214738	35989	75130
Rate per 100000 population	51.3	11515.2	443.7	19.0	144.5	106.5

	DATASET 3			DATASET 4		
	Cluster 1	**Cluster 2**	**Cluster 3**	**Cluster 1**	**Cluster 2**	**Cluster 3**
Centroid	414500, 418750	461800, 495700	436950, 493540	414500, 418750	461800, 495700	436950, 493540
Radius	5km	5km	5km	20km	20km	20km
Cases	141	32	153	228	69	132
Population	148069	165	18255	1214738	35989	75130
Rate per 100000 population	95.2	19393.9	838.1	18.8	191.7	175.7

Table 2.39: Location of clusters using Chris Brunsdon's datasets.

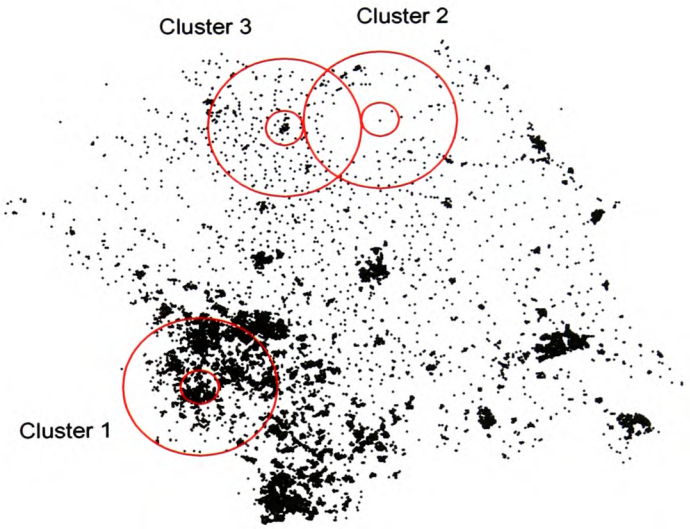


Figure 2.17: Location of clusters.

2.9.2.1. Turnbull's method v Kulldorff's spatial scan statistic

Various population sizes were used for Turnbull's method to identify the most likely clusters. The corresponding population proportions were used for Kulldorff's local spatial scan statistic to enable a comparison to be made. Table 2.40 and table 2.41 show

the analysis for varying population sizes for Turnbull's method and Kulldorff's method respectively.

	DATASET 1		DATASET 2	
Population	**p-value**	**centroid**	**p-value**	**centroid**
500	0.001	461800, 495700	0.053	437420, 493120
20000	0.001	461800, 495700 (464500, 498300)	0.001	453900, 505900 (453100, 508200)
40000	0.001	453900, 505900	0.001	461800, 495700
80000	0.001	449700, 500500	0.001	443200, 499960
150000	0.001	444200, 507600*	0.001	446900, 501100

	DATASET 3		DATASET 4	
Population	**p-value**	**centroid**	**p-value**	**centroid**
500	0.001	461800, 495700	0.001	442200, 480400
20000	0.001	461800, 495700	0.001	436300, 485100
40000	0.001	454800, 506600 (454700, 506200)	0.001	442900, 489500
80000	0.001	449500, 503100	0.001	446800, 496900
150000	0.001	446500, 489800*	0.001	451900, 490600

** 5 other most likely clusters exist, all very close to each other.*

Dataset 1 and dataset 2 (population=20000) and dataset 3 (population=400 produced two most likely clusters of equal significance.

Table 2.40: Cluster analysis using Turnbull's method.

	DATASET 1			DATASET 2		
Population	**p-value**	**centroid**	**radius**	**p-value**	**centroid**	**radius**
0.010%	0.001	461800, 495700	8.4km	0.201	437420, 493120	0.0km
0.415%	0.001	461800, 495700	13.0km	0.001	449500, 503100	10.1km
0.830%	0.001	453900, 505900	21.6km	0.001	461800, 495700	21.0km
1.660%	0.001	454700, 506200	25.5km	0.001	443200, 499960	24.2km
3.112%	0.001	454800, 506600	25.8km	0.001	446900, 501100	35.3km

	DATASET 3			DATASET 4		
Population	**p-value**	**centroid**	**radius**	**p-value**	**centroid**	**radius**
0.010%	0.001	464900, 498300	6.2km	0.002	442200, 480400	0.0km
0.415%	0.001	461800, 495700	14.0km	0.001	436300, 485100	7.9km
0.830%	0.001	454800, 506600	22.4km	0.001	442900, 489500	13.2km
1.660%	0.001	453900, 505900	25.4km	0.001	446800, 496900	24.6km
3.112%	0.001	453900, 505900	25.4km	0.001	451900, 490600	32.4km

Table 2.41: Cluster analysis using Kulldorff's method.

Cases highlighted in red show where the same cluster centroid was identified for both methods (note that both methods identified a non-significant cluster using dataset 2 for the smallest population size due to the actual cluster in this area being much larger and the population size used was too small to identify clusters of 20km in geographical size). Both methods produced the same centroid as a cluster for all population sizes using dataset 4. Both methods identified the main cluster in dataset 1 for small population sizes. However, for larger population sizes both methods tended to located one large cluster rather than two smaller clusters close to each other. This was true for all datasets, especially dataset 2 and dataset 4 where the clusters overlapped. Both methods failed to identify two separate clusters, although the clusters obtained occurred in the correct area. Thus, to summarise, both methods are dependant on the maximum population size entered to locate possible clusters. These produced most likely clusters in the correct area of the study region for one of the actual clusters but of varying radii (comparing these results with the actual locations in figure 2.17).

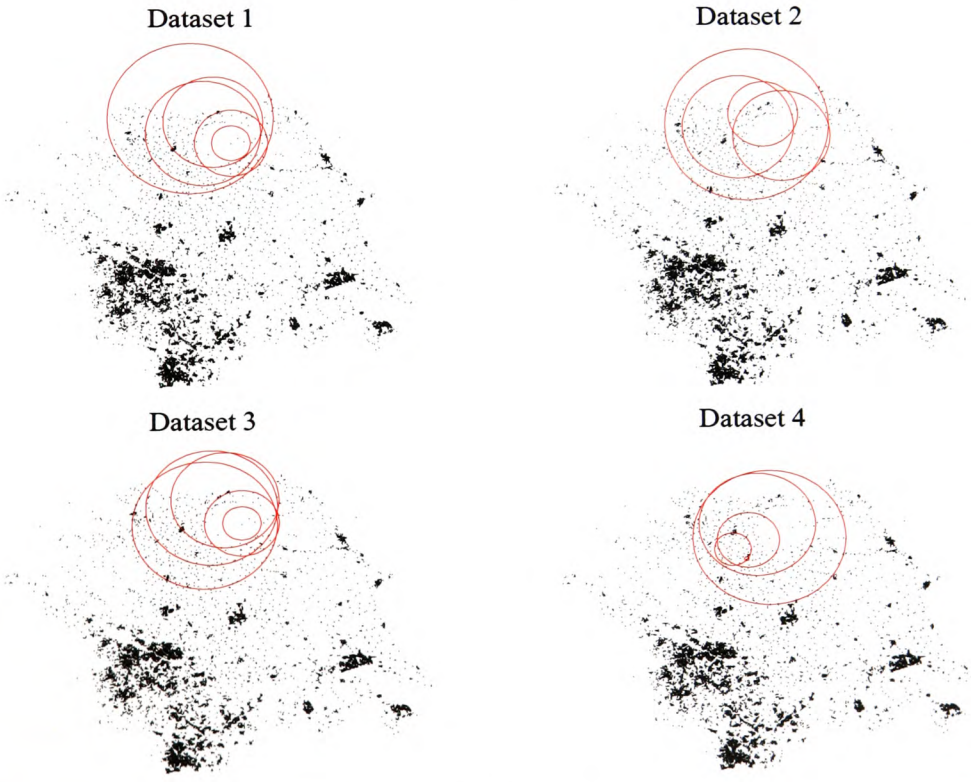


Figure 2.18: Locations of the most likely cluster using Turnbull's method.

Figure 2.18 and figure 2.19 show the locations of the most likely cluster for both methods for various population sizes. It is clear which cluster is for which population size due to the increasing size of the cluster radii (note that dataset 2 and dataset 4 only show four clusters since the smallest population size identified a cluster of just 1 ward, hence a radius of 0km).

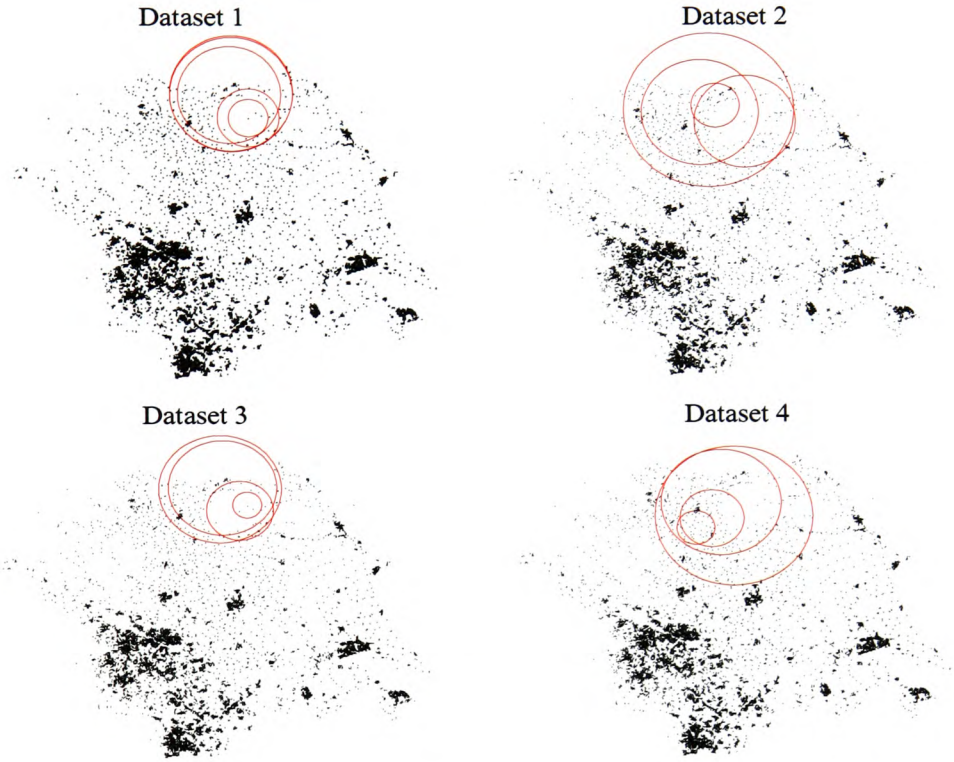


Figure 2.19: Locations of the most likely cluster using Kulldorff's method.

To summarise, both methods appeared to locate the same most likely clusters for varying population sizes. However, there appeared to be a problem in identifying clusters when such clusters were close to each other. Since cluster 2 and cluster 3 of radius 20km in dataset 2 and dataset 4 overlapped and no geographic overlap was selected, SaTScan V5.1.3 was unable to locate both clusters and only produced one larger cluster in this region. Note that clusters do not have to be a given shape or homogeneous in risk. In reality, clusters will not be circular in shape. For simplicity, the clustering algorithms

assumed that the clusters were circular. More recently, Kulldorff has defined a clustering algorithm that searches for clusters shaped as ellipses (Kulldorff et al., 2006).

Both Turnbull's method and Kulldorff's method had problems in locating the correct sized cluster although Kulldorff did have the advantage in that it displayed clusters with no geographical overlap. The other advantage that Kulldorff has is that it can adjust for confounding in the data (e.g. sex, age, social class – although this has not been investigated here), unlike Turnbull's method in ClusterSeer V2.2.4. Thus, if two clusters overlapped each other, Kulldorff would fail to identify them if no geographic overlap was selected and tends to produce one cluster with cases observed in both clusters.

2.9.2.2. Kulldorff's spatial scan statistic v Score Test of Lawson and Waller

Table 2.42 shows the analysis of Kulldorff's method for the four datasets and for a maximum population size of 0.5% and a maximum population size of 5%. The actual foci for each of the datasets shown in table 2.39 were used as the centroid of any cluster.

		Maximum population size = 0.5%			Maximum population size = 5%		
		414500,418750	461800,495700	436950,493540	414500,418750	461800,495700	436950,493540
Dataset 1	**radius**	1.664km	12.954km	7.624km	6.385km	26.396km	35.630km
	p-value	0.001	0.001	0.001	0.001	0.001	0.001
	cases	14	91	92	95	204	232
Dataset 2	**radius**	1.168km	16.243km	1.448km	1.168km	41.338km	27.587km
	p-value	0.629	0.001	0.001	0.830	0.001	0.001
	cases	4	33	16	4	173	129
Dataset 3	**radius**	1.753km	13.984km	8.417km	6.475km	26.553km	35.630km
	p-value	0.001	0.001	0.001	0.001	0.001	0.001
	cases	26	177	176	175	380	405
Dataset 4	**radius**	No cluster identified	16.243km	8.417km	3.100km	40.353km	35.085km
	p-value		0.001	0.001	0.894	0.001	0.001
	cases		43	43	74	280	240

Table 2.42: Kulldorff's spatial scan analysis.

Table 2.42 shows that the size of the cluster identified depends on the maximum population size that was used. Only dataset 2 produced the same sized cluster for both

population sizes (cluster 1). However, from table 2.42 this focus contained the rate of 19.0 per 100,000 population within 20km and was therefore not an actual cluster – it was one of the two "clusters" that contained a lower rate. However, the radius of the cluster found in table 2.42 is of 1.168km containing 4 observed cases but with a non-significant p-value. Thus, a smaller "cluster" was found but Kulldorff correctly identified this as being non-significant. Comparing the radii in table 2.42 with the actual clusters in table 2.39, it seems that for dataset 1 and dataset 3, a smaller maximum population size should have been used since the radii were over 5km for cluster 2 and cluster 3 as in the real clustered dataset. For dataset 2 and dataset 4, it appeared that a maximum population size between 0.5% and 5% was required since the radii for these were lower and higher than 20km respectively. Thus, prior knowledge of the population distribution within this area could have helped in the identification of cluster 2 and cluster 3 using Kulldorff's method. Generally, the actual clusters have not been found using Kulldorff's method due to the values entered for the maximum population size found in the clusters. Also, in dataset 2 and dataset 4, two of the clusters overlapped (both of radius 20km) but when using this method, no geographic overlap of clusters was selected.

Table 2.43 shows the p-values obtained for the four datasets along with the three actual focused clusters when using the Score Test of Lawson and Waller. The tests were run for three different radii around the foci to explore the stability of the p-values obtained. Dataset 2 and dataset 4 show similar p-values for distances within 5km and 20km of the foci. The focus 414500, 418750 contained the "lowest rate" cluster in each of the datasets, but produced non-significant results for this focus even though the rate was much higher than the background rate. Looking back at the rates inside the actual cluster, table 2.39 dataset 3 has the highest rate; hence this result was "more significant" than the other three datasets. The Score test correctly identified a significant cluster for the foci 461800, 495700 and 436950, 493540 in dataset 01 and dataset 03 when using the whole dataset but could not identify them when using smaller distances from the foci. It is clear from table 2.43 that the p-values of 0.001 are highly significant and have a very high rate inside the cluster. For dataset 2 and dataset 4, the actual clusters were 20km in radius and overlapped each other; hence this could be the reason for the p-values obtained when

using the whole dataset. i.e. the rates inside these clusters were very similar to each other and since they overlapped, the results were not as significant as one would have initially thought. However, figure 2.20 shows the cumulative observed and expected numbers of cases within the two overlapping foci within 20km of the centroids (using the whole dataset). It can clearly be seen that the excess number of observed cases to the expected cases does not occur until approximately 12km from the focus 461800, 495700 and hence the non-significant result.

Distance	Focus	Dataset 1	Dataset 2	Dataset 3	Dataset 4
Within 5km	**414500, 418750**	0.644	0.571	0.287	0.566
	461800, 495700	0.466	0.508	0.462	0.461
	436950, 493540	0.470	0.516	0.477	0.459
Within 20km	**414500, 418750**	0.560	0.552	0.098	0.569
	461800, 495700	0.688	0.513	0.810	0.484
	436950, 493540	0.743	0.505	0.844	0.452
Whole dataset	**414500, 418750**	0.579	0.554	0.076	0.559
	461800, 495700	0.001	0.505	0.001	0.004
	436950, 493540	0.002	0.021	0.001	0.024

Table 2.43: Score test analysis.

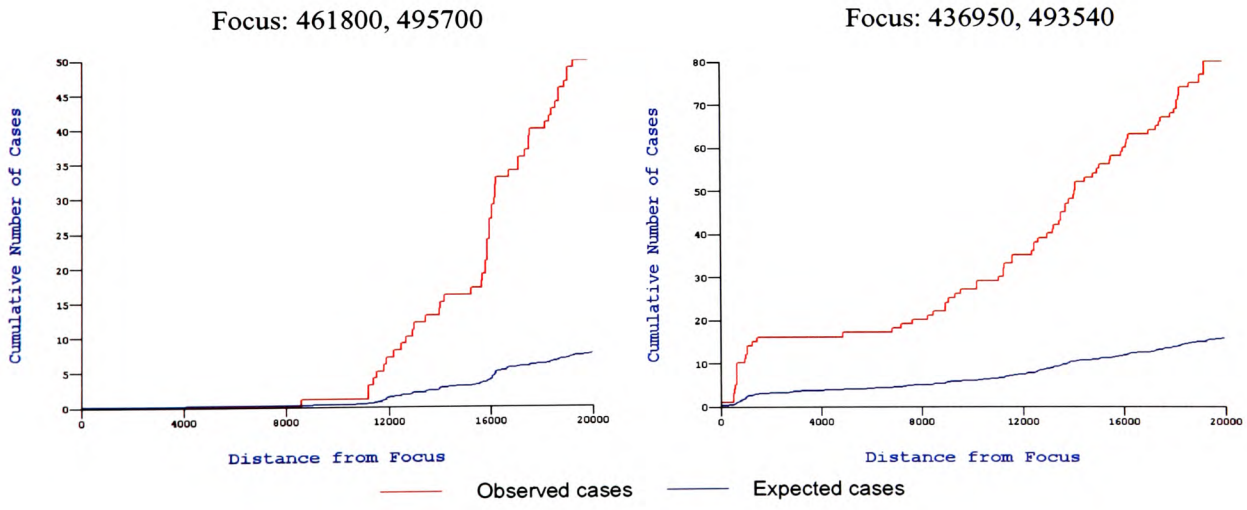


Figure 2.20: Dataset 2 (whole dataset) analysis.

As with Kulldorff's analysis, due to dataset 2 and dataset 4 having clusters of radius 20km that overlapped each other, highly significant results were expected due to the rates found inside these clusters. However, the p-values obtained were not as significant as expected. Thus care must be taken if two or more clusters are known to exist very close together. Also, it is not known how the risk varies within the "cluster" using the Score test – it could be uniform.

To summarise, it appears that Kulldorff's method has advantages over Turnbull's method in that it produces no overlapping clusters, thus locating more than one cluster area and the size of the cluster is accurate unless the clusters are very close to each other (Turnbull is also poor in this case). Kulldorff's method should also be used over the Score test since it has the advantage that the maximum population size can vary and still locate the actual cluster; however if clusters are located near to each other or are overlapping, the method tends to locate one larger cluster. Kulldorff's method was able to find the located clusters in Openshaw's datasets (as a local method and a focused method) but failed to correctly identify the clusters in Brunsdon's datasets. Turnbull's method produced clusters of varying observed cases depending on the population size entered for Openshaw's datasets and Brunsdon's datasets. The Score test aided in providing the cumulative observed and expected cases from a specific focus but did not give an overall cluster analysis within a specific radius. However, it should be noted that these eight datasets all showed very high rates inside the clusters. It would be worthwhile obtaining many other simulated datasets with varying degrees of clustering to determine if these clustering algorithms produced consistent results.

2.10. Conclusions

Analysis of the leukaemia cancer dataset and simulated datasets have identified the spatial scan statistic as being the most robust of the algorithms examined in terms of correctly identifying a cluster, especially when the population size was varied. Although many algorithms produced similar results, there were advantages and disadvantages associated with each algorithm. There are likely to be different situations when one or more algorithms are applicable depending on particular factors.

When determining if a cluster exists, confounding is an important factor that should be taken into account. Only a few algorithms could allow this in ClusterSeer V2.2.4 so was not accounted for in the case and population at risk data.

In general, care must be taken when a single test is run and a result of borderline significance is obtained. If a p-value of borderline significance is obtained it is advised that the test should be run a number of times to calculate the mean p-value. It should also be noted that if different time periods were used e.g. four year periods or ten year periods then the p-values obtained and clusters obtained would have given different results. For example, a cluster found in a five year period may disappear in a ten year period if hardly any cases were located in the cluster region in the following five years.

Global methods

Moran's I statistic does not identify locations of clusters; this method is a test for departures from spatial randomness and not an actual cluster test. However, the disadvantage to this method is the inability to take population at risk into consideration. There were 2,835,141 persons aggregated into 908 wards using the 1991 Census. The 908 wards vary in population size from 423 to 16,965 with a mean of 3122 and a median of 2254. The lower and upper quartiles were 1499 and 4002 respectively indicating a right skewed distribution. It is clear that population at risk is not constant between wards and the number of cases per ward will not take this into consideration. Therefore, population at risk should be taken into account when using the geographical boundaries of wards. Other geographical boundaries may be suitable to run this method depending on the distribution of the population at risk for the boundaries used.

Oden's I pop method is an extension of Moran's I statistic but is able to take population at risk into account. Thus, if population at risk data are available, Oden's I pop method should be used over Moran's I method. As with Moran's I method it does not locate clusters, it detects departures from spatial randomness and provides the user with the

proportion of the test statistic that is attributable to clustering in the dataset and whether neighbouring wards are similar in excess.

Sensitivity testing of the Besag and Newell method showed that this method displayed erratic results when increasing the size of a possible cluster by just one case for one of the periods studied. Thus due to these unstable results, the Besag and Newell method was disregarded.

Cuzick and Edwards' method should be used as a global method if using case and control data (also allows for confounding) since this was the only available method to use in ClusterSeer V2.2.4 for case and control data. Oden's I Pop method should be used if case and population at risk data are available. Note that these methods do not locate areas of clustering, only if clustering is present in the datasets.

Local methods

When analysing the local methods with the leukaemia dataset, Besag and Newell's method identified all possible locations of clusters. Due to the large number of clusters identified, it was very difficult to visualise the clusters obtained for the Besag and Newell method; although this method did provide the user with the general location of clusters. Turnbull's method identified the three most likely clusters; however each cluster often contained part of another cluster that was also obtained. i.e. geographic overlay was present. Turnbull's method had the advantage over other methods in that if the population at risk of a perceived cluster was known, this could be used as the maximum population size to identify whether the cluster was significant or not (if one did actually exist).

Anselin's local Moran test is related to Moran's I test and although it does not take population at risk into consideration it located areal units where neighbouring rates were dissimilar to neighbouring geographical units and identified this location as an outlier. This method is advantageous in that it locates outliers so that the user can explore these

anomalies in more detail. e.g. identifying neighbouring population figures to determine whether the outlier is "geographically different" to its neighbours.

Turnbull's method and Kulldorff's local spatial scan statistic were compared using 9999 Monte Carlo randomisations. Turnbull's method was run 50 times to compare the p-values obtained, Kulldorff was run only once since the pseudo-random number generator that was used results in the same p-value for each run. The advantage that Kulldorff has over Turnbull's method is that none of the clusters overlapped whereas the method by Turnbull usually generated clusters that overlapped. The actual cluster was located in the correct area irrespective of the size of the population entered, unlike Turnbull's method.

Focused tests

Two focused methods used to locate clusters in Wales were the Score Test of Lawson and Waller and Bithell's linear risk score test and should be used when population at risk data are available. Both methods provided similar results. Bithell's linear risk score had the additional option to choose one function from four to model the observed data and to also choose initial values for two parameters. Although non-significant results were found for this dataset, results of borderline significance for other datasets should be treated with caution since changing the risk function used could change the outcome from a significant result to a non-significant result. Depending on the choice of relative risk function, variable p-values were produced using Bithell's method. A cumulative plot of observed and expected cases from the focus is useful since the user can identify other areas that should be investigated, if any arise throughout the study region. However, this can be done without any clustering algorithm.

Another focused method analysed was Diggle's method and was again subject to two parameters to use a raised density model. Varying p-values were obtained for different values of the two parameters which again could cause the user to determine no clustering in a dataset when in fact clustering was present if p-values were of borderline

significance. This was found to be the case for small populations at risk. i.e. areas in Mid Wales as opposed to populated areas in North and South Wales.

Kulldorff's method was used as a focused method and like previously, it had the advantage of no overlapping clusters present. However, no clusters were located within 20km of NYG landfill site using population at risk data or controls.

To summarise, the Score test of Lawson and Waller gives useful information such as the observed and expected plots but does not show the size of the cluster– the p-value supplied is the overall p-value for the dataset that is entered into the model. Kulldorff's method should be used over Diggle's method if case and control data are available. Even though Diggle's method produced similar p-values when analysing data around NYG landfill site (large population), irrespective of the choice of parameters, this method produced varying p-values when a much smaller population was analysed (e.g. Trawsfynydd in North West Wales). Hence, Kulldorff's method should be used with case and control data.

Case-control methods

Two methods in ClusterSeer V2.2.4 were case-control methods; those being Cuzick and Edwards' method and Diggle's method. Although they do not use population data, they use a method by which controls are selected to reflect the population distribution throughout Wales as a whole. However as stated earlier, the Cuzick and Edwards' method does not locate the clusters. Diggle's method is a focused test and thus can only locate clustering and calculate significance from a point source; it is only suitable to use when the location of a perceived cluster is known. To use Diggle's method as a local clustering method would entail each ward centroid to be entered as a potential cluster and run 908 times (for Welsh wards). However, other methods described previously were more suitable to use as a local clustering method rather than using Diggle's method.

Space-time methods

Most of the space-time clustering methods available in ClusterSeer use case data only. These were disregarded in the initial stages of this analysis and hence the only space-time algorithm that could be analysed was Kulldorff's space-time scan statistic. Thus, this method is the only one that can be used for locating clusters in a specific time frame.

Simulated datasets

When looking at simulated datasets of Openshaw, the remaining algorithms generally performed well with the "no clustering" dataset. If any "clusters" were located, p-values of over 0.9 were obtained for many methods indicating no clustering. The Score test of Lawson and Waller showed no excess of observed cases or expected cases in the dataset. For the other datasets, Kulldorff's method generally performed the best since it located the actual cluster areas and tended to identify the actual cluster size. Turnbull's method generally increased the cluster size area for increasing population sizes whereas Kulldorff was not sensitive to this issue.

Examining Brunsdon's datasets, cluster 2 and cluster 3 were very close to each other in dataset 2 and dataset 4 and overlapped in area. Most algorithms tended to locate one larger cluster in the area as opposed to two separate clusters. Kulldorff's Scan Statistic also had trouble in locating two clusters for most population sizes. However, for very small population sizes at less than 0.5% of the total population, Kulldorff's method identified the cluster that consisted of the highest rate, cluster 2 for dataset 1 and dataset 3. From the results in section 2.9 it is suggested that Kulldorff's spatial scan statistic be used over Turnbull's method due to no geographical overlapping of clusters and the stability of results irrespective of the choice of maximum population size. Kulldorff's focused cluster test should be used over the Score test since it performed better with Openshaw's datasets and it did locate the correct sized cluster for small population values. As stated previously, the rates found inside the clusters in the simulated datasets were very high compared with the background rate and thus, in reality, the clustering

algorithm should have located the majority of the clusters. It is advised that other simulated datasets be used to analyse further these clustering algorithms.

As stated earlier, a large simulation exercise is required to evaluate the clustering algorithms. Only eight simulated datasets were evaluated here for various clustering algorithms. However, the spatial scan statistic appears to produce consistent results when identifying clusters. Circular clusters were located in these datasets for simplicity, but in reality, clusters may be other shapes rather than circular.

Most appropriate algorithm(s)

To summarise, the free downloadable software SaTScan V5.1.3, the Kulldorff Scan Statistic appeared to be the most effective algorithm used when analysing the datasets. This method had the advantage in that it could analyse both case data with population at risk data and case data with control data. In everyday life, population at risk data or control data may be unavailable; it was beneficial for this research in that both methods could be compared. Although it has many user defined options which can give alternative results due to the values given, many of the other algorithms in ClusterSeer V2.2.4 cannot overcome this. The algorithm produced the "most" correct results when analysing the simulated datasets by Openshaw. A disadvantage was that if clusters were very close together as was the case with two of Brunsdon's datasets, the user defined option of "no geographical overlap" tended to dilute the two clusters that were near to one another into one overall cluster. Although this option was not available in ClusterSeer V2.2.4, the other methods tended to produce too many clusters since they took every ward centroid as a possible cluster so that those methods that provided graphics could not be interpreted. Another advantage to this method was that it could be used as a local clustering method, a space-time clustering method and a focused clustering method. Thus most types of clustering are possible to investigate when using SaTScan V5.1.3. Also, Kulldorff's method allowed the user to take confounding into account whereas most methods in ClusterSeer V2.2.4 did not allow the user to take this major factor into account.

Oden's I Pop method should also be used as a global clustering method when population at risk data are available and Cuzick and Edwards' method should be used when control data are available. It should also be noted that Anselin's local Moran test is a beneficial tool in that it identifies outliers; thus this method should be used to assess the dataset for any irregularities which can then be explored in more depth. However, this method should not be used as a clustering method, only as an aid to investigate outliers.

3. THEME TWO

Determining the population at risk around hazardous sources

3.1. Aims and objectives

It is important to identify possible increased risks associated with sources of pollution. These sources can take many forms, but this area of work concentrates on landfill sites and electric power lines. The former are examples of point sources where an exposed area at risk is typically within a specific radius from the point source. These are generally assumed to be circular in shape for simplicity, but, in the real world can be non circular depending on the direction of exposure. For electric power lines, which are linear, an exposed area at risk is typically within a specified distance from the power line for simplicity, but, like other hazardous sources, can be defined according to other factors such as wind speed and topography.

There were several objectives to be tackled in this theme. These objectives were evaluated by analysing datasets of leukaemia, childhood cancer and brain cancer, in relation to landfill sites and electric power lines:

- To describe in detail the main methods of aggregating population at risk for any exposed area.
- To compare the results of these methods to determine the extent to which resulting conclusions agree or disagree. Extrapolation of population, definition of the "unexposed" population at risk and choice of geographical unit are explored to determine if these influence results.
- To determine whether an increased risk of cancer exists around landfill sites and electric power lines in Wales.
- To investigate the effect of latency periods for two clusters of brain cancer and leukaemia.
- To examine the comparability of results of methods with the spatial scan statistic when taking the latency period into account.

3.2. Background

Previous studies have expressed public health concerns around landfill sites and electric power lines (Fielder et al., 2000, Redfearn et al., 2002, Winterfeldt et al. 2004). There are many hundreds of landfill sites that have been, or are still, in operation throughout Wales. Landfill sites can be thought of as a point source, defined by coordinates (eastings and northings). Electric power lines (over 132kV) are linear sources. Figure 3.1 shows the difference between a point source and a linear source.

The top left hand corner of figure 3.1 shows a landfill site located by a red triangle. The coordinates can relate to the central position of the landfill site or the opening gate of the landfill sites or some other location within the landfill site. An area within a specified distance, in this case of radius 2km, has also been defined around the landfill site. This is also known as a buffer. The population at risk exposed to the landfill site within this area has been shaded. Thus, any person living in the shaded area is assumed to be exposed to the landfill site.

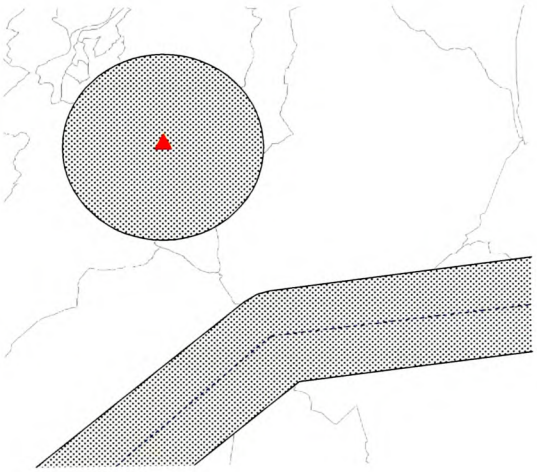


Figure 3.1: Point source and linear source.

The lower right hand corner of figure 3.1 shows a linear source, such as an electric power line, represented by a blue dashed line along with a buffer of 1km either side of the linear source. The shaded area represents the population at risk exposed to the linear source.

Electric power line theory

Electrical currents generate electric and magnetic fields, collectively called electromagnetic fields (EMFs). Electric and magnetic fields occur naturally; examples of electrical fields are from thunderstorms and magnetic fields deep inside the molten core of Earth. Frequency defines how fast the field oscillations are occurring and is measured in Hertz (Hz). In the UK, electricity systems produce fields of around 50 Hz whereas in North America they are approximately 60 Hz.

Electric fields are produced by voltage. Voltage is the pressure behind the flow of electricity. In UK homes these electric fields are approximately 230 volts, but outside they can be distributed from 11,000 volts to 400,000 volts. The higher the voltage, the stronger the electric field. The strength of electric fields is measured in volts per metre (V/m). Electric fields cannot easily penetrate buildings, hedges or fences.

Magnetic fields are produced by current. Current is the flow of electricity and is measured in Amperes (Amps). The higher the current, the stronger the magnetic field. Magnetic fields are measured in microteslas (μT). There are 1000μT in 1 millitesla and 1.25μT in 1 Amp per metre.

In 2002, the UK government issued guidelines for exposure levels to the general public – these are 12,000 volts per metre for electric fields and 1600μT for magnetic fields (ICF consulting, 2002). The duration of exposure to electric and magnetic fields is also an important factor.

All overhead power lines produce fields. The fields are greatest directly under the lines and fall rapidly with distance to the sides of the lines. In theory they could reach around 100μT at ground level but in general are much lower than this. Table 3.1 shows maximum and typical field strengths for various pylons (Energy Networks Association, 2007).

Type of power line	Maximum or typical field level	Microteslas μT	Volts per metre V/m
Largest steel pylons (275kV and 400kV)	Maximum field (under line)	100	11000
	Typical field (under line)	5 - 10	3000 - 5000
	Typical field (25m to side)	1 - 2	200 - 500
Smaller steel pylons (132kV)	Maximum field (under line)	40	4000
	Typical field (under line)	0.5 - 2	1000 - 2000
	Typical field (25m to side)	0.05 - 0.2	100 - 200
Wooden poles (11kV and 33kV)	Maximum field (under line)	7	700
	Typical field (under line)	0.2 - 0.5	200
	Typical field (25m to side)	0.01 - 0.05	10 - 20

Table 3.1: Maximum and typical field strengths for pylons.

Electric fields are at typical ambient levels at approximately a distance of 100m from a 40,000 volt power line. Magnetic fields reach typical ambient levels between 200m and 300m from high voltage power lines (ICF Consulting, 2002). Electrical substations can produce magnetic fields of up to 2μT at the substation and very often no electric field at all. A few metres away from the substation, the fields will be indistinguishable from other fields experienced at home (Energy Networks Solutions, 2004). Due to their construction, underground cables produce no external electric field.

In the UK, averaged over a 24-hour period, magnetic fields experienced are less than 0.2μT (Energy Networks Associations, 2007). In only 0.5% of UK homes do magnetic fields exceed 0.4μT.

At work, power station workers experience only a few μT on average during working hours. An electrician will typically experience 1μT and an office worker approximately 0.3μT (Energy Networks Association, 2007).

3.3. Literature Review

The following review is split into three sections. The first section details previous techniques used in studies to aggregate the population at risk and the number of observed cases, the second section relates to studies regarding landfill sites and the third section relates to past studies regarding electric power lines and childhood cancers or adult cancers. Potential biases are detailed along with other problems in order to try to explain why they may give differing results or misinterpretation of results. Table 3.2 shows a summary of past studies regarding landfill sites and electric power lines.

Methodologies of past studies

Briggs et al. (2007) stated that census data were essential to provide population denominators for estimating rates of disease to determine those exposed. However, problems occurred when determining intercensus year population denominators at small area level due to the changing geographical units that have been created between censuses.

Geographic zones were defined in studies by Openshaw (1981, 1984) to represent population data but depending on the definition of the zone, results vary. This is the reason as to why many studies tend to use administrative boundary data as census data are official.

In the past, the Small Area Statistics Health Unit (SAHSU) defined the population at risk by aggregating wards or enumeration districts within the specified area of interest (Aylin et al. (1999), Elliott et al. (1992), Morris et al. (2003)). Post 2001, SAHSU tended to use super output areas and output areas which were the 'building blocks' of the 2001 census. However, the problem of obtaining population data at small area level for inter-census years as described by Briggs et al. (2007) was apparent when analysing data over a number of years. Presently, SAHSU use postcode data which has recently become available to define the population at risk (exposed). However, the problem with using

postcode data are that no age breakdown is available so assumptions regarding the age distribution have to be made.

Dunn et al. (2001) compared three different methodological approaches to analysing a spatially referenced dataset regarding the potential health effects near to a wallpaper factory. The first method involved the 'traditional' epidemiological approach whereby a control area was selected based on the similarity of socio-economic status, size, proximity and qualitative similarity to the study area. This resulted in the analysis of nine enumeration districts in the control area and eight enumeration districts in the study area. Odds ratios were calculated with multiple logistic regression to adjust for confounding variables. The second approach used GIS software to calculate prevalence rates for three concentric bands 0-500 metres, 500-1000 metres and 1000-1500 metres from the study area as a function of estimated exposure to factory emissions. Multiple logistic regression was used to test if there was a linear relationship between reported ill health and distance from the factory, accounting for confounding variables such as age, sex and smoking. An additional part of the GIS analysis involved air quality modelling. A mathematical air quality model (Gaussian plume model) was used to define the sectors to be analysed. Finally, raised incidence modelling was used to determine the extent of clustering of cases around the point source. Essentially, a ratio is calculated based on that the odds that a person at a particular point is a case as opposed to a control. However, this depends on factors such as the overall prevalence of cases in the population, other risk factors (covariates) and a decay function from the point source. Results showed that there were some inconsistencies in the results which would have meant a different conclusion depending on the method used. It was found that the results from the GIS and epidemiological analysis agreed less than those from either of the other two combinations. This variation is not unlikely since although each method addresses the same question, different models are used to analyse the data. In general however, it was concluded that the results complemented rather than contradicted or duplicated each other.

A few years later, Dunn et al. (2007) used the GIS approach and raised incidence modelling approach described above for the incidence of Legionnaires' disease in relation to proximity of residence to cooling towers. The GIS approach used two concentric circles around each cooling tower of distances 0-500 metres and 500-1000 metres. These were then split into four sectors for each of the concentric circles. Population data from the 1981 census were used to derive population data for enumeration districts by five year age group and sex. Directly standardised rates and indirectly standardised ratios were calculated for the eight sectors. Those living within 1000 metres were used as the reference category. The raised incidence modelling compared the cases of Legionnaires' disease with a control population of lung cancer. Again, results were complementary rather than contradictory since although the same question was addressed, different techniques were used in the analysis. Thus, some methods as seen in this study by Dunn can clarify results from another method whereas other methods such as those by Briggs et al. (2007) when using postcode data or address data can contradict results.

Past studies relating to landfill sites

There have been many previous studies regarding landfill sites and their effects on congenital malformations (Palmer et al. (2005), Vrijheid (2000), Dolk et al. (1998)) but very few on the association between landfill sites and cancer. The few studies that are available are examined here along with potential biases in each study.

All cancer deaths before a child's 16th birthday for the period 1953-1980 were analysed by Knox (2000) in relation to 460 toxic-waste landfill sites in England and Wales, along with 377 incinerators (307 hospital incinerators and 70 municipal incinerators). Distances were compared from source to birth address, and from source to death address for those children that had moved house. However, no significant results were found when examining the solid-waste landfill sites in relation to childhood cancer (ratios were calculated depending on those migrations where either one or both addresses were within a particular radius from a landfill site, or those migrations with one address inside and the

other address outside a specified radius from a landfill site. The most extreme results obtained were a ratio of 1.04 for one or both addresses being within 3km of a landfill site and a ratio of 0.92 when one address was inside and one address was outside a radius of 5km from a landfill site). Other combinations were also explored but no significant results were found. An element of recall bias was evident since the interviewees were the child's parents and details regarding specific information about the cancer may not be correct. Various other factors may have influenced the results, such as confounding or latency (the period from exposure to diagnosis) not taken into account, and lack of information relating to the exposed cases. Various cancers were analysed but multiple testing issues were not mentioned.

In a Welsh study, WCISU (2001) were asked by RANT (Residents Against Nant-Y-Gwyddon Tip), a pressure group, to investigate a possible increase in non-Hodgkin's lymphoma around the Nant-Y-Gwyddon landfill site in South Wales from 1988 to 2001. WCISU concluded that there was a marginal significant increase in non-Hodgkin's lymphoma within 2.5km of the landfill site for the later four year period 1998-2001 compared with the corresponding rates in all Wales. Other landfill sites in Wales were investigated to determine whether the risk at Nant-Y-Gwyddon was greater than others, however no significant results were found. Ongoing surveillance of non-Hodgkin's lymphoma in this area has since shown a decrease in relative risk for the period 1998-2001 due to other non-Hodgkin's lymphoma cases being diagnosed in other parts of Wales for this time period and hence the relative risk decreasing slightly in this area from significant to borderline. The Agency for Toxic Substances and Disease Registry (ATSDR) was asked to take part in a review of the Nant-Y-Gwyddon independent investigator's recommendations in 2002. The ATSDR reported that "the epidemiological evidence does not provide much support for a relationship between exposures to the site and all causes of mortality; mortality from specific causes; the incidence of cancers; or the rates of adverse reproductive outcomes such as birth defects, low birth weight, and spontaneous abortions". In the study by WCISU, latency was taken into account as a crude measure by only allowing landfill sites that had been operating for at least five years by the study period, 1983-2001 in the analysis. Further investigation of the effect

of latency is warranted. There was also a lack of exposure detail known on the cases regarding residential history, exposure to particular gases and occupational history of cases. Selection bias was evident (post-hoc analysis) since a pressure group approached WCISU claiming of an increased risk in this area and subsequent work was carried out with the knowledge of this supposed effect. These factors could have altered the final conclusion, especially for the final four year period 1998-2001 which was of borderline significance.

Postcode data were used in a study in Great Britain by Jarup et al (2002) who analysed various cancers diagnosed in the period 1982-1997 and living within 2km of 9565 landfill sites in Great Britain that had operated at some time during the study period. Adjustments by age, sex, year of diagnosis, region and deprivation were taken into account for this study. The population at risk in the exposed area was determined by postcodes that are located within 2km of each of the landfill sites. Those postcodes that lay outside the 2km buffers were classed as the reference category (not exposed). The population at risk was adjusted for intercensus years. To allow for latency of cancer, a crude adjustment of analysing data for the period 1987-1997 (5 year latency period) was used for adult cancers and 1983-1997 (1 year latency period) for childhood cancers. Six cancer types were studied. However, no significant increased risks of cancers such as bladder, brain, hepatobiliary cancer or leukaemia were found near landfill sites. Confidence intervals were narrow for most cancers due to the very high number of cases in the exposed category. Results showed that there was evidence that the populations living near landfill sites were more deprived compared with the population in the reference category for unexposed. The landfill sites may have been subject to error in terms of location, operating dates and classification of waste. Postcode locations were only accurate to 100 metres. Migrational effects were noted but were probably less of an effect here due to the large area of the "exposed" population due to the number of landfill sites analysed. This study did not take into account whether the case was actually exposed to the landfill site in terms of the site's operation period, it only took into account whether the site was in operation at the time of the analysis period. This could

affect the results since a large proportion of the population at risk would no longer have been exposed at the time of diagnosis.

Other studies outside the UK include the New York State Department of Health (NYSDOH, 1998) who analysed cancer incidence within 38 landfill sites in New York State for the period 1980-1989. Distances investigated were 250 feet (33 landfill sites), 500 feet (4 landfill sites) and 1000 feet (1 landfill site) from landfill sites depending on the dispersion of landfill gas that was identified. Cancers examined included brain, bladder, kidney, liver, lung, leukaemia and non-Hodgkin's lymphoma. Controls were identified as a random sample that did not have cancer and resided within the zip codes that were used in the analysis. A statistically significant fourfold increase for female bladder cancer and female leukaemia was found. The study reports a lack of detail on the data regarding type of exposure and lack of duration of exposure and confounding could have played a part into the significantly increased risks. Thus, there is definite evidence of lack of exposure details, confounding not taken into account and evidence of multiple testing issues which should have been accounted for.

Goldberg et al (1999) investigated the distance from a municipal solid waste landfill site in Montreal, Canada for thirteen cancer sites for males, ages 35 years to 70 years for the period 1979-1985. Logistic regression was used to calculate odds ratios and adjusted for covariates such as age, cigarette consumption, ethnicity and body mass index. Ethical approval was obtained to enable face-to-face interviews with respondents. Response rates varied from 8% from 3730 respondents (or surrogates) (i.e. cases) to 69% from 533 subjects from a population-based control group (i.e. controls). Recall bias was evident regarding some of the confounding variables. Additionally, the numbers of cases in the exposure categories were very small. The cases and controls were classified into "high", "medium" and "low" geographic zones depending on their distance to the landfill site. Various distance categories were used to determine significance of results. Multiple testing (e.g. analysing various cancers, different distances, various years of diagnosis) should be taken into account. Various multiple testing techniques are summarised in section 2.2. For this analysis, calculating 99% confidence intervals in place of 95%

confidence intervals would help to overcome this issue. An adjusted odds ratio of 2.0 was calculated for living within 1km of the landfill site for non-Hodgkin's lymphoma; the 95% confidence interval (CI) (1.0, 4.0) indicates a borderline significant result. As with many of the previous studies, recall bias, small numbers and multiple testing issues are evident here.

Past studies relating to childhood cancers and electric power lines

Most studies have analysed childhood leukaemia in relation to electric and magnetic fields. Other studies tend to analyse all childhood cancers. Childhood cancer studies have the advantage over adult cancer studies in that children are less likely to have moved home from an exposed area to an unexposed area or vice versa and thus, place of residence is a good proxy for exposure. Also, children tend to be at home more than adults. The major disadvantage with studies of childhood cancers is that the number of observed cases is usually very small, especially when breaking down these cancers into further subtypes. Most studies define an exposure over $0.4\mu T$ to be high and assess this measurement for possible increased risk – however, it can be very difficult to measure an individual's magnetic field exposure. However, many studies generally have very few cases in this category. Thus confidence intervals are generally very wide indicating the large degree of uncertainty. An increase of just one or two more cases in this exposure category could dramatically affect the relative risks and change a significant result to a non-significant result or vice versa.

Twenty five years ago, a study by Wertheimer and Leeper (1979) suggested a link between residential exposure to extremely low frequency magnetic fields (ELFMF) and childhood cancer. This was the first published study of its kind. Since then many other published studies have produced variable and contradictory results regarding associations between exposure to ELFMF and cancer. Most of these studies involve childhood cancer, leukaemia, brain cancer, tumours of the central nervous system and female breast cancer.

A study by London et al (1991) analysed childhood leukaemia in children aged under 11 years, using 232 cases from a population based tumour registry and 232 controls obtained through friends and random digit dialling. Measurements of magnetic field in the child's bedroom, spot measurements of magnetic and electric fields and wiring configuration were used to define exposure categories. A significantly increased risk was found for very high current relative to very low current and underground configuration combined with an odds ratio of 2.15 (95% CI: 1.08 to 4.28). Studies like this, where controls are found from random digit dialling are biased since they only sample homes with telephones. People of very low socioeconomic status are harder to reach by this method and are thus underrepresented; however this will only matter if socioeconomic status is a confounder and so depends on whether or not the cancer incidence is related to it. In summary, there is evidence of selection bias, confounding variables not taken into account and very small numbers in this study.

Two years later in 1993, 140 cases of childhood cancer (defined as ages 0-19 years) were observed in a study by Verkasalo et al. within 500m of overhead power lines for the period 1970-1989 in Finland in a cohort study. Various childhood cancers were studied. Indirect standardisation was used to calculate standardised incidence ratios. The only statistically significant increased risk was for tumours of the nervous system (males) for magnetic fields greater than or equal to 0.2μT. This study does not allow for multiple testing even though various cancer sites have been tested - a significant result will occur by chance for every 1 in 20 tests. The main disadvantage of this study is the lack of robust estimates due to the small numbers in the relevant exposed categories. Assumptions were made regarding the magnetic field strength experienced by patients since their day to day movement was not known.

In order to achieve high statistical power, a large epidemiological study was conducted by the United Kingdom Childhood Cancer Study Investigators (UKCSSI, 1999) of 3636 cases of childhood cancer diagnosed in the United Kingdom. A four year period was studied. This case-control study used controls matched by sex, date of birth (month and

year) and region of residence (2 controls per case). For the magnetic field analysis, only 2226 cases had EMF measurements and were thus entered into the analysis, resulting in the very small numbers in the highest exposure categories. Five sources were used to measure EMF exposures: child's home, overhead power lines nearby, external-source questionnaire, electric appliances at home questionnaire and measurements in schools. The EMF component of the study used only one control of the two controls selected earlier. A significant increased risk of central nervous system tumours was found for magnetic fields between 0.1μT and 0.2μT, although this was based on just 25 cases and 10 controls. No other significant results were found, even for magnetic fields over 0.4 μT. Even in a large study of this magnitude, small numbers are still an issue in the highest exposure category producing uncertain results due to the width of confidence intervals. There was also evidence of selection bias since there was under-representation of individuals living in more deprived areas among controls compared with cases.

One year later in 2000, the UKCCSI published a study that took into account effects due to exposure to chemicals, to ionising radiation, and the possibility of abnormal responses to infections. A population based case-control study covering England, Wales and Scotland - 3338 cases and 7629 controls was analysed. Controls were matched by sex, date of birth and region of residence. Measurements were taken by detectors in houses of the extent of radon gas, terrestrial gamma radiation and electric and magnetic fields. There was no association found between any malignancy and measured power frequency magnetic fields. This study showed evidence of participation bias since 64% of control families were interviewed compared with 87% of case families. A socio-economic difference was observed whereby controls who were deprived were less likely to participate than controls who were more affluent. An additional study by UKCCSI in 2002 found no significantly increased risks based on data for the period 1992 to 1996 with magnetic field measurements taken from the children's bedroom and living room based on 273 cases and 276 controls.

The study by UKCSSI above was incorporated into a pooled analysis of magnetic fields and childhood leukaemia by Ahlbom et al (2000) who stated that future studies would

only be of use if selection bias and confounding could be adequately addressed. It also advised that there should be sufficient numbers of cases and controls in the highest exposure category. The pooled analysis showed a significantly increased risk for measurement studies (actual measurements made in people's homes) in the category >=0.4μT but a non-significant increased risk for calculated field studies (population based data are used) – combining both types of studies resulted in a significant increased risk for this exposure level with 44 observed cases.

Measurement studies (those studies where information is available regarding the individual's home) can be affected by selection bias, since controls with a low socio-economic status are less likely to participate than cases with a low socio-economic status as was the case by the UKCSSI (2000). Population registries were used for many of the calculated field studies eliminating possible selection bias. However, a disadvantage of calculated field studies is that they do not take into account characteristics of individual's homes. Wartenburg (2001) specifies selection bias, information bias and confounding as the three main factors that influence studies towards biased results. These factors are evident in these studies.

The most recent childhood study by Draper et al (2005) from the Childhood Cancer Research Group (CCRG) studied 29,081 childhood cancers diagnosed in the period 1962-1995 in England, Scotland and Wales of which 9700 were of leukaemia. Controls were matched by sex, approximate date of birth and birth registration district. Eight distance measures from electric power lines were used, those being 0-49m, 50-99m, 100-199m, 200-299m, 300-399m, 400-499m and 500-599m and over 600m. Results showed that children living within 600m had a significantly increased risk of leukaemia compared with those that lived over 600m from electric power lines. No other significant results were found for other childhood cancers. The study quotes that the significant results could have been due to chance or by confounding. Confounding factors such as deprivation were not taken into account which could have influenced the results. These results do not fit within general theory regarding large distances from electric power lines stated earlier.

Past studies relating to adult cancers and electric power lines

Most studies of adult cancers in relation to magnetic fields study cancers such as leukaemia and brain cancer.

Coleman et al (1989) investigated leukaemia and residence near electricity transmission equipment, defining exposure to be those living within 100m of an overhead power line or substation. Each case was matched with a control by age, sex, year of diagnosis and district of residence and registered with a solid tumour excluding lymphoma. Other distances were also investigated. Of 771 leukaemia diagnoses between 1965 and 1980 (resident in four adjacent London boroughs), 84 cases were aged less than 18 years. For adults, the relative risk of leukaemia relative to cancer controls for residence within 100 metres of an electric power line was 1.45 (95% CI: 0.54 to 3.88).

Another case-control study in the United Kingdom by Youngson et al (1991) analysed a much larger number of cancer patients, 3144 diagnoses of adult haematological malignancies between 1983 and 1985 in the North West and Yorkshire Regional Health Authorities (of which only 7% lived within the vicinity of a power line in the study design). For each case, a single control matched to the case by age, sex, year of diagnosis and health district of residence, was chosen from inpatient hospital discharges – an element of selection bias was evident due to an "ill" control population. Five distance categories were assessed; less than 25m, 25m to 50m, 50m to 75m, 75m to 100m and greater than 100m. An odds ratio of 1.29 (95% CI: 0.99-1.68) was calculated in Yorkshire for living less than 50m from overhead power lines (compared with greater than 100m), a borderline significant result. Magnetic field analysis produced no significant results but, in common with most of these studies, few cases were found in the higher categories of magnetic field exposure. A statistically significant result was shown for myeloid leukaemia between 50m and 75m, OR=2.88 (95% CI: 1.22-6.82) but no overall trend with distance was found.

Outside the UK, Feychting et al (1994) analysed adult cancers aged 16 years and over, with exposure defined as those living between 0 and 300m from a 220kV or 400kV power line for the period 1960-1985 in Sweden. Data were obtained from the cancer registry and 325 cases were included in the study. The familiar problem as with previous studies was the small numbers of cases in exposure categories. The results showed no significantly increased risks. Only 7 cases of chronic lymphocytic leukaemia (CLL) had exposure greater than or equal to 0.2μT. A further study by Feychting et al (1998) analysed 699 female and 9 male breast cancers over the period 1960-1985. One control for each case was matched by age, sex, parish and residence near the same power line. Again no significantly increased risks were found. Only 2 male breast cancer cases had exposure greater than or equal to 0.2μT.

A larger distance again was used in a study by Verkasalo et al (1996). A cohort of 383,700 persons (2.5 million person-years in total), aged 20 years and over, diagnosed between 1970 and 1989 in Finland was analysed. Cases of leukaemia were compared to matched controls living within 500m of overhead power lines in the period 1974-1989 – 203 cases in total, on average approximately 12 cases per year which is very small for such a study. Current, distance and typical locations of phase conductors were taken into account. No significant results were shown. However as in the childhood studies, the total number of cases in the highest exposure category was low (5 cases between 1μT and 2μT, and 4 cases greater than or equal to 2μT). This study also included 1229 cases of female breast cancer. However no significantly increased risks were found, a standardised incidence ratio (SIR) of 0.75 was calculated with a 95% CI of 0.48 to 1.10 for cumulative exposure greater than or equal to 0.2μT.

Li et al (1997) analysed 870 pathologically confirmed cases of leukaemia from Northern Taiwan for the period 1987-1992. One control per case, matched on date of birth, sex and date of diagnosis, was used in the analysis. Average and maximum magnetic fields using distance from the power lines were calculated along with height of wires above the ground. The risk of adult leukaemia among those exposed to magnetic fields of more than 0.2μT was calculated relative to the risk among those exposed to magnetic fields of

less than $0.1\mu T$. Again, as with many of the previous studies with few cases, only 3 cases of chronic lymphocytic leukaemia were analysed for magnetic fields greater than $0.2\mu T$. A slightly elevated risk was found for all leukaemia for magnetic fields greater than $0.2\mu T$, odds ratio 1.4 (95% CI: 1.0-1.9) and acute lymphocytic leukaemia (ALL) greater then $0.2\mu T$, odds ratio 1.7 (95% CI: 1.0-3.1) based on 17 cases.

Most of the above studies tend to have a source of bias, be it selection of controls or recalling events. However, Kheifets (2001, pS128) suggests that "biases that may be present in some of the individual studies appear to be countervailing and are unlikely to have a substantial influence on the overall estimate of the relative risk."

Comparisons of the above studies in terms of their results are very difficult due to the varying distances, diseases and years of diagnosis studied. However, many of the studies do not take confounding into account, have small numbers in exposure categories, selection bias of controls and lack of detail known on exposed cases which will all influence the resulting conclusions. Past studies have varied in their definition of "childhood" cancer. Some studies only analyse those cases under 11 years of age while other studies have analysed those cases under 21 years of age. However, age variation is very unlikely to have caused such biased results.

Vrijheid (2000) has reviewed previous studies of the epidemiological impacts of landfill sites including cancer incidence and found that such studies are affected by confounding factors (the main one being socioeconomic status not taken into account), methodological problems and potential biases, and that more multisite landfill studies are required to enable a large population at risk to determine whether an increased risk exists. Such studies have tended to show differing conclusions regarding increased or decreased risks around landfill sites.

In summary, the following problems tend to occur with past studies:

- Selection bias/Recall bias
- Confounding not taken into account

- Lack of detail on actual exposures
- Short follow-up times
- Small numbers
- Multiple testing
- Healthy worker effect
- Latency periods unaccounted for
- Estimates of population at risk
- Methodologies adopted
- Statistical techniques used
- Temporal problems
- Exposure Assessment

This area of research should overcome the problems of small numbers, latency periods, population at risk estimates and methodology adopted.

Such studies have highlighted a number of methodological concerns which hinder transferability of findings to other contexts. Very few studies have investigated the implications of employing different techniques of population estimates when calculating population at risk in relation to landfill sites. A number of potential reasons for this situation can be suggested; for example, it is often difficult to obtain the exact population of interest within the analysis area due to the lack of data or their expense. This theme suggests that the methods of population estimation available could be an important determinant on findings and that there is a clear need for sensitivity analysis using a range of population estimation techniques. We illustrate this with reference to a relatively new data set that has become available to researchers in the UK context. This is described in more detail in the next section. Previous studies tend to only have postcode accuracy to 100 metres. It was clear from the first theme that the ward that a case is placed in can change depending on the accuracy of the postcode. Therefore, the enumeration district change will be even greater. The analysis presented here uses postcode accuracy to one metre resolution.

LANDFILL SITES				
Author	**Year**	**Cancer**	**Method**	**Results**
Goldberg et al	1995	All Cancers	Case control study. Proximity to solid waste site and wind direction	Increase in incidence of stomach, liver, lung and prostate for men, stomach and cervix-uteri for women
Williams et al	1997	All Cancers	Case control study. Cases living within 3km of waste disposal site, 1974-1991.	SIR=380 (95% CI: 139-827) for male brain, 2 other significant clusters for periods within the time period for female breast and uterine cancer
Jarup et al	2002	All Cancers	Population based study. Residence within 2km of a landfill site, 1982-1997	No significant results for landfill sites
Knox et al	2000	Cancer Deaths, <16 years.	Case study. 22458 childhood cancer deaths, 1953-1980. Comparing distances from suspect sources to the birth address and to death address.	Migrations where either 1 or both addresses were within 3km of a landfill site, ratio =1.04, non signfiicant
WCISU	2001	NHL	Population based study. Cases within 2.5km of NantYGwyddon landfill site in Wales 1983-2001.	RR=1.70 (95% CI: 1.12-2.60) for 1998-2001
ELECTRIC POWER LINES				
Author	**Year**	**Cancer**	**Method**	**Results**
Wertheimer and Leeper	1979	Childhood Cancer, 0-19 years	Case control study, 344 cases and controls allocated to type of wire code	No risk estimates presented
London et al	1991	Childhood Leukaemia	Case control study, 211 cases & 205 controls allcoated to type of wire code	OR=2.2 (95% CI: 1.1-4.3) for very high current configuration wire code (42 cases)
Verkasalo et al	1993	Childhood Cancers	Population based study, 140 cases, 1970-1989. Calculated historical and cumulative magnetic fields.	No significant results
UKCSSI	1999	Childhood Cancers	Case control study (2 controls per case). 3838 cases. 1992-1996	OR=2.4 (95%CI: 1.2-5.1) between 0.2 and 0.4 μT (central nervous system tumours)
UKCSSI	2000	Childhood Cancers	Case control study (2 controls per case). 33838 cases. 1992-1996	No risk estimates presented
UKCSSI	2002	Childhood Cancers	Case control study (1 control per case). 426 cases. 1992-1996	OR=1.32 (95%CI: 0.73-2.39) for all leukaemia, OR=0.90 (95%CI: 0.59-1.35) for all malignancies
Draper et al	2005	Childhood Cancer, 0-14 years	Case control study. 29081 children diagnosed with cancer in England and Wales, 1962-1965. Distance from home address at birth to nearest high voltage overhead power line.	Within 200m, OR=1.69, 95%CI:1.13-2.53. Within 200m-600m, OR=1.23, 95%CI:1.02-1.49) compared with over 600m
Feychting et al	1994	All Cancers >16 years	Case control study, 1960-1985. Analysis within 300m of a power line in Sweden.	AML: RR=1.7 (95% CI:0.8-3.5), CML: RR=1.7(95% CI:0.7-3.8) for >=0.2 μT
Coleman et al	1989	Leukaemia	Case control study, 771 cases. Distance and magnetic field strength were analysed.	RR=1.45 (95%CI: 0.54-3.88) <=100m, RR=2.0 (95%CI: 0.4-9.0)<=50m compared with >100m
Juutilainen et al	1990	Leukaemia	Male industrial workers, 1971-1980. Unadjusted for confounding.	RR=1.4 (95%CI: 1.1-1.8) for possible exposure, RR=1.9 (95% CI: 1.0-3.5) for probable exposure
Youngson et al	1991	NHL, Leukaemia	Case control study (1 control per case), 3144 cases. 1983-1985. Distance and magnetic field analysis.	OR=1.29 (95%CI: 0.99-1.68) for <50m of a power line
Verkasalo et al	1996	Leukaemia>=20 years	Population based study, 1974-1989. 203 cases within 500m of a power line were analysed. Current, distance and typical locations of phase conductors were taken into account.	SIR=0.71(95%CI: 0.19-1.81) for >=2μT (based on 4 cases)
Li et al	1997	Leukaemia	Case control analysis (1 control per case) 1987-1992. Average and maximum magnetic fields using distance from power lines was calculated.	OR=1.4 (95% CI: 1.0-1.9) for >0.2μT (all leukaemia)
Preston-Martin et al	1996	Central Nervous System tumours, 0-19 years	Case control analysis. Mean magnetic fields were calculated.	OR=2.3 (95% CI: 1.2-4.3) for underground cables

Table 3.2: Literature Review of various studies.

3.4. Datasets

Landfill site data were obtained from the Environment Agency for this area of research. WCISU were given data for all landfill sites in Wales that were over 25,000 cubic metres in capacity that had taken non-inert waste and had operated at some point since 1973. This dataset was reduced to all sites greater than 500,000 cubic metres in volume, to concentrate on all large landfill sites in Wales, and had operated at some point during the period 1982-2001 (the period covered by the cancer datasets for analysis). These criteria gave 77 landfill sites in Wales (36 of which were greater than one million cubic metres in volume). National grid references (eastings and northings) in this dataset were obtained for the central position of each landfill site (or as close to the centre as possible).

Electric power line data were provided by Dr Mary G Wright of Bristol Oncology Centre at The University of Bristol. The pylon grid references were interpolated at the University of Glamorgan to give line coordinates for electric power lines greater than or equal to 132kV in Wales. As a means of validating the overhead power lines data supplied by Dr Wright, a GIS layer of power lines in Wales was supplied by Western Power Distribution for power lines greater than or equal to 132kV which covered the South Wales area. Additionally, a GIS layer was obtained from National Grid Transco of the electricity transmission system of England and Wales as of June 2005 that contained all 275kV and 400kV overhead electric power lines. The original file from the University of Bristol was subsequently used for all analysis due to the agreement between all files.

For electric power line analysis, childhood cancer and leukaemia were investigated using previously used methods in other studies to determine whether an increased risk existed. Each of the cancer datasets was obtained from the WCISU for the twenty year period 1982-2001 at individual level including age at diagnosis, sex, address, postcode, eastings and northings of postcode (correct to 100 metres), enumeration district using 1991 Census and quintile of deprivation using the Townsend score for Welsh enumeration districts from the 1991 Census. Population figures used in calculations were from the

1991 Census. The population figures were stratified by sex and five year age band for all enumeration districts and wards in Wales and aggregated to the number of years being analysed. These were obtained from Census Area Statistics on the Web (CASWEB). Two geographical units were used: enumeration districts and wards. The Townsend quintile of deprivation for each geographical unit enabled the aggregation of population by five year age band, sex and quintile of deprivation. Table 3.3 shows the total number of cases of the cancers studied and those cases that could not be analysed due to invalid postcodes or no allocation to a Townsend quintile of deprivation.

Cancer	Total	Not included	% Not included	Analysis
Brain and CNS Tumours	9535	138	1.4	9397
Leukaemia	7715	127	1.6	7588
Childhood Cancer	1370	0	0.0	1370

Table 3.3: Cases included in analysis.

Table 3.4 and figure 3.2 show the total number of cases by Townsend quintile of deprivation (from affluent, Q1, to deprived, Q5) and corresponding crude rates per 100,000 population for the cancers examined. Table 3.4 also shows the total person years at risk in Wales (to the nearest thousand) for the period 1982-2001 for all ages and for those aged 0-14 years for males and females. Table 3.5 shows a summary of the rates per 100,000 population for each of the cancer sites studied in Wales for the period 1982-2001.

MALES					
	Q1	**Q2**	**Q3**	**Q4**	**Q5**
Brain Cancer	975	887	924	950	916
Leukaemia	839	825	829	920	855
Childhood	159	134	140	137	166
Person-years	5357000	5062000	5275000	5600000	5973000
Child-years, 0-14	1037000	946000	1004000	1093000	1478000
FEMALES					
	Q1	**Q2**	**Q3**	**Q4**	**Q5**
Brain Cancer	938	921	890	1033	963
Leukaemia	594	634	677	719	696
Childhood	119	137	100	113	165
Person-years	5606000	5380000	5662000	6082000	6507000
Child-years, 0-14	992000	892600	958000	1036000	1407000

Table 3.4: Numbers of cases by ED Townsend quintile of deprivation.

Type of Cancer	Rate*
Brain Cancer	16.6
Leukaemia	13.4
Childhood Cancer, 0-14 years	12.6

** Rate per 100,000 population for the twenty year period*

Table 3.5: Crude rates for various cancer sites in Wales, 1982-2001.

There were 9397 cases of tumours of the brain and central nervous system (ICD 10 codes C701-C729, C732-C734, D220-D229) analysed for the period 1982-2001 in Wales, of which 49.5% were males. There appears to be a decreasing trend in crude rates per 100,000 population for males (blue) from affluent to deprived which can be seen in figure 3.2. The same pattern is not seen in females (red).

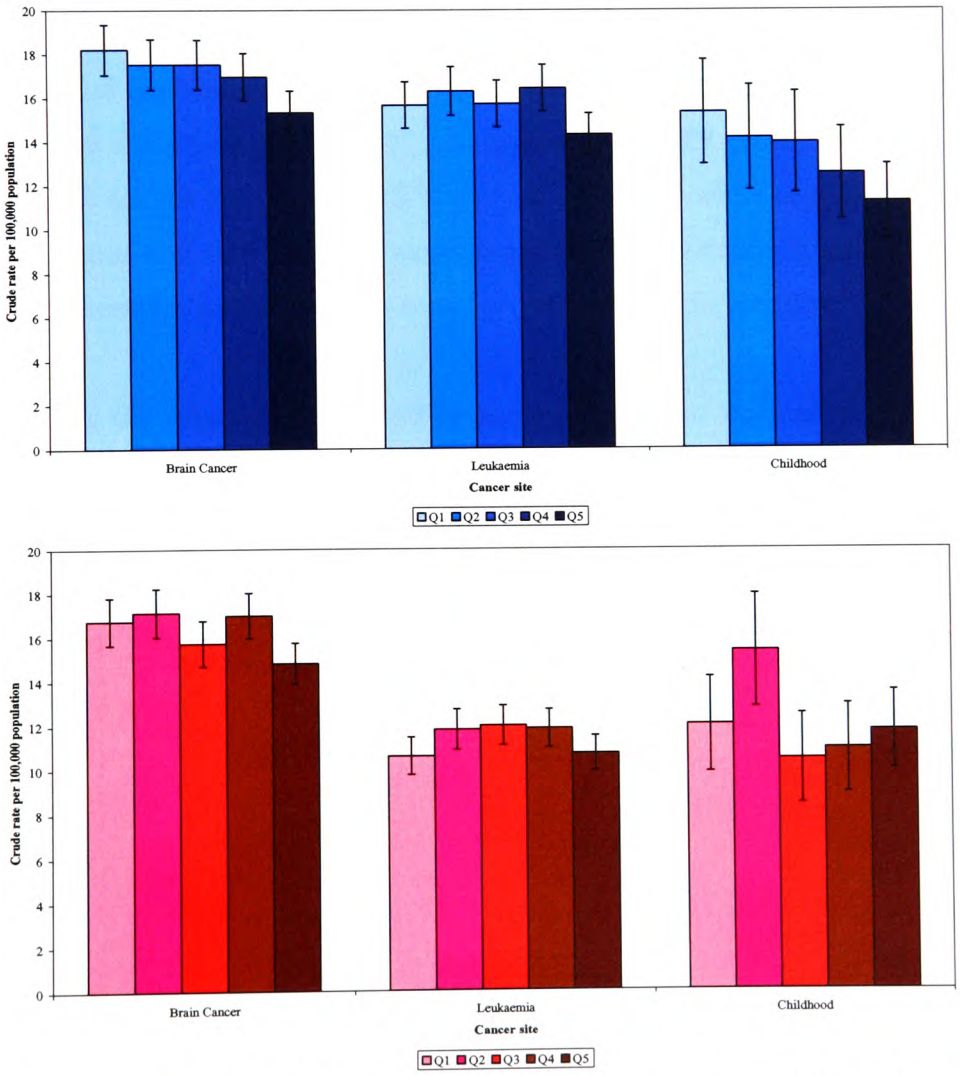


Figure 3.2: Crude rates per 100,000 population by Townsend quintile of deprivation.

7588 cases of leukaemia (ICD 9 codes 2040-2089, ICD 10 codes C910-C959) were analysed for the period 1982-2001 in Wales, of which 56.2% were males. Crude rates per 100,000 population for females show the central quintile (marginally) as having the highest crude rate whereas quintile four has the highest crude rate for males. Note that there are a slightly lower number of leukaemia cases than in the analysis concerning the first theme regarding clustering algorithms. This is due to the fact that when the data for the other cancer sites were extracted, a new dataset for leukaemia was also extracted. Cancer registration is a dynamic process and figures are updated on an ongoing basis, especially for later years. Cases can be amended, added or deleted at the WCISU due to reasons such as misclassification of diagnosis or late registrations from hospitals.

All cases of childhood cancer (ICD 9 codes 1400-2089 excluding 1730-1739, ICD 10 codes C000-C969 excluding C440-C449 and ages less than 15 years) were extracted from the WCISU database for the period 1982-2001. 1370 cases were analysed, of which 736 (53.7%) were males and 634 (46.3%) were females. A decreasing trend from affluent to deprived can be seen for males but this trend is not apparent in females.

Figure 3.3 shows the crude rates per 100,000 population for the cancers studied by sex and by year of diagnosis and show a generally increasing trend for all cancers examined.

Figure 3.3: Crude rates per 100,000 population in Wales, 1982-2001.

3.5. Methodology

The research in this project extends analysis that WCISU has previously conducted (for one landfill site) to an all Wales study using a new method proposed by the Small Area Health Statistics Unit (SAHSU) to analyse such data. The SAHSU is situated at Imperial College in London and was set up in 1987 following the Black enquiry (Black, 1984) regarding the incidence of leukaemia in children and young adults near the Sellafield nuclear power plant. SAHSU estimate the effect of an environmental factor on specified diseases in relation to the population at risk and generally define exposure as a circular area at a specified distance from the point source. Multiple landfill sites in Wales are examined in terms of the risk associated with them, along with pre- and post-opening site analysis. Clustering algorithms are used to aid in the development and interpretation of the identification of clusters near hazards in Wales based on the residential postcodes at time of diagnosis. Later, latency periods are explored. Since this work dealt with patient specific data, the majority of work was carried out at the WCISU due to confidentiality issues.

Elliott et al (2001) quote that 80% of the British population live within 2km of a landfill site. This distance is usually used by the Small Area Health Statistics Unit (SAHSU) for analysing such data around the point source; they tend to analyse inner (less than 2 kilometres) and outer bands (usually 2 kilometres to 7.5 kilometres) from the point source. Thus, to be consistent, a distance of 2km was used as the "exposed" population around a landfill site to determine whether an increased risk existed for the cases residing within 2km of landfill sites in this area of research. The cancer datasets analysed regarding landfill sites were brain and central nervous system tumours and leukaemia, which were studied by Jarup et al in 2002.

There are various geography levels that can be used in the UK. However, a suitable geography level is required whereby population data are available. Census information contains the most accurate information. Thus, geography levels from the 1991 Census or 2001 Census should be used. Geographical boundaries were obtained from the website

UKBorders to obtain wards and enumeration districts (ED) from the 1991 UK Census (the mid-point of the twenty year dataset). The population figures for wards and EDs in Wales used to calculate rates and expected figures were obtained from CASWEB from the 1991 Census. An assumption for this analysis was the aggregation by year for ward and ED population. The datasets used were for a twenty year period, thus the population figures by ward and ED were multiplied by 20 to obtain the total person years at risk. If new housing estates were created in wards or EDs then the population in those wards or EDs would be much larger than the figures in the analysis provided here. Population adjustment is sometimes used to take into account the annual change in population at ward or ED level since only Census data are available (1991). The effect of this is explored later.

3.6. Description of methods

Three methods were used to estimate the population at risk to determine whether an increased risk exists within 2km of a landfill site; these methods were the intersection method, centroid method and SAHSU's postcode method.

Assume an area of interest is defined as the exposed region – this is known as the buffer and can be defined as any shaped region. For illustration (and simplicity) purposes for landfill sites, the buffer was defined to be the area within a radius r, of the landfill sites. The geographical units, for example EDs, inside the buffer are noted. These geographical units are used to build up the area of the buffer in order to estimate the population at risk. If an entire ED is wholly contained within r then the population at risk is known for that ED using census data. Problems occur if the ED is partly contained in the circle. How is the population classified as exposed or not exposed? One technique is to allocate an ED to be included in the exposed area or not based on whether or not its centroid (either the geometric centre of the ED or a population weighted centroid) is within the buffer. Another technique is to aggregate the population at risk if any part of the ED is within the buffer. A postcode method used by SAHSU, aggregates the population at risk in the exposed area by using a percentage of the population based on the percentage of domestic

houses inside the buffer at risk compared with all domestic houses in that particular ED for those EDs that intersect the buffer. These three methods are detailed further.

3.6.1. Intersection method

This method included the cases and population at risk in any ward or ED that was contained wholly within the exposed area, along with any ward or ED that intersected the buffer. The major disadvantage of using this method is that risk estimates are biased due to a proportion of population within particular geographical units included in the analysis that were not actually "exposed".

For information, figure 3.4 shows the landfill sites in Wales (triangles) and wards and EDs highlighted grey whose ward or ED was contained within 2km of a landfill site for the intersection method. To put this into context, an area of approximately 12.5km^2 is examined around each landfill site in Wales. It can be seen that the areas highlighted at risk are clearly of different sizes.

3.6.2. Population weighted centroids method

This method identified centroids of a particular geography level (e.g. wards or EDs) within distance of the point source in question. All geographical units whose centroids were contained within the buffer of radius *r* were classed as exposed even if part of the geographical unit was contained outside the buffer. Geographic centroids (geographic centre of a particular ward) or population centroids (the centroid is weighted towards the population distribution within that ward) were available but only population centroids were used here.

Thus, for landfill sites, all wards or EDs whose population weighted centroid was within 2km of a landfill site were included in the analysis. Similarly, this method contained cases and population data inside the buffer that were not included in the analysis if the ward or ED centroid was contained outside the buffer, and cases and population data outside the buffer were included in the analysis as exposed. If a true high risk did exist within a specific point source then the effect would be diluted.

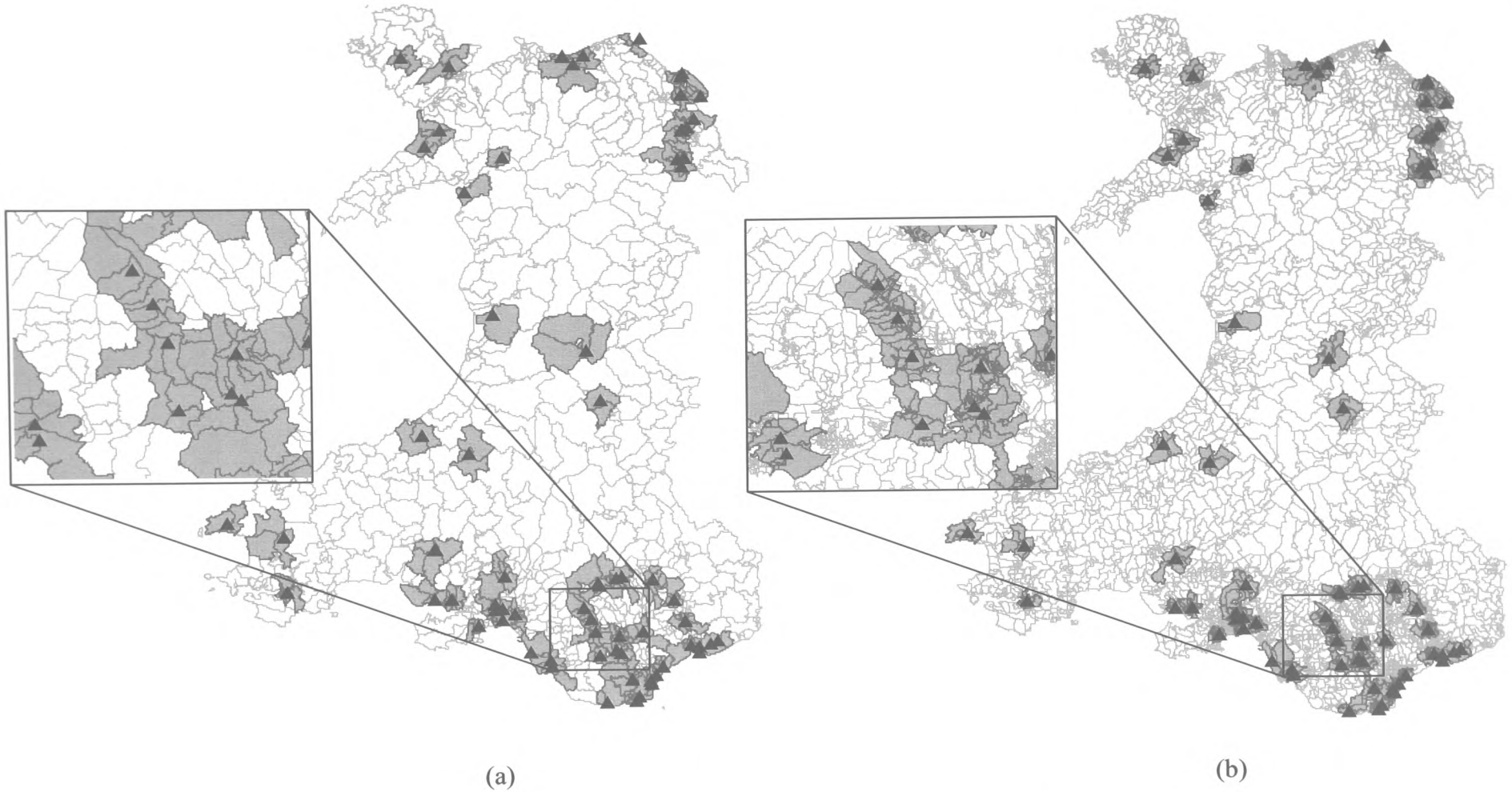


Figure 3.4: "Large" landfill sites in Wales operational between 1982 and 2001 and highlighted wards (a) and EDs (b) that lie within 2km of a landfill site using the intersection method.

3.6.3. Postcode method

The following method adopts a new approach whereby the Ordnance Survey product CodePoint™ is used. CodePoint™ is a database of all current postcodes in the United Kingdom. It contains information for each postcode such as the easting and northing to one metre resolution (the average of all houses with that postcode), thus enabling the highest accuracy possible, total number of domestic delivery points, ward identification and other information.

In summary, to calculate the number of exposed cases and the population at risk using the postcode method, the following steps were taken:

Number of cases

- The postcode for each case in the dataset was identified.
- The postcode at diagnosis was updated to its current postcode using the extension ProAddress in ArcGIS. Those cases not able to be updated to a current postcode were excluded from the analysis.
- CodePoint™ was used to obtain the current postcode to one metre resolution (the coordinates of each postcode is the average position of the number of houses with a particular postcode).
- Those cases whose postcode centroid was within the buffer were identified as those exposed.

Population at risk

- CodePoint™ was used to obtain all current postcodes in Wales correct to 1 metre resolution (average of all houses with a particular postcode).
- All geographical units (wards or EDs) that intersected the buffer were identified since those wards or EDs contained wholly inside the buffer contribute all the population at risk for that particular ward or ED.
- For each geographical unit that intersected the buffer, all postcodes were identified whose centroid was:

(i) contained within the geographical unit.

(ii) contained within the buffer.

- Those postcodes identified in the previous action were used with CodePoint™ to calculate the total number of all domestic delivery points:

  (i) contained within each geographical unit.

  (ii) contained within the buffer.

- The proportion of each geographical unit whose delivery points were within the buffer (for all those delivery points in each ED) was calculated for those geographical units that intersected the buffer.
- For those geographical units that intersected the buffer, the proportions of those delivery points within the buffer for each geographical unit was applied to the corresponding population figures by sex and five year age band from the 1991 Census to determine the population at risk for those geographical units.

This method is similar to the centroid method in that postcodes were classed as exposed if their centroid is within the buffer, but the advantage is that postcodes are smaller than wards or EDs. However, the population per postcode is not known in CodePoint™. The number of delivery points for each postcode is known so those delivery points inside the buffer as a percentage of all delivery points is applied to the population of each ward or ED to estimate the population at risk for that particular ward or ED that intersects the buffer.

Assumptions that were made for this method were as follows:

- The distribution of people in each household for those in the exposed population at risk was the same as for those not exposed.
- No postcode had been terminated and then reinstated at a different place (this should not be the case).
- Cases that could not be assigned a current postcode were excluded from the analysis (this applied to less than 0.5% of cases that had a postcode at diagnosis – some cases did not have a postcode for diagnoses in early 80s – see table 3.4 for numbers of cases excluded – these were distributed randomly throughout Wales).

- Over the twenty year period no additional houses were built or destroyed with an existing postcode since this could cause the postcode centroid to move position and may or may not move outside the buffer.

This method was used for wards and EDs to compare the results. Results should be similar since the exact same area was analysed but slightly different population figures were used, depending on the proportion of the ward or ED that was included in the analysis.

The method presented here improves the accuracy of the base population at risk for the area of interest and thus the analysis of cancer within 2km of landfill sites. A comparison is made with two traditionally used techniques to show the differences between results. Figure 3.5 explains the process of analysing the area of interest in diagrammatic form to calculate the population at risk.

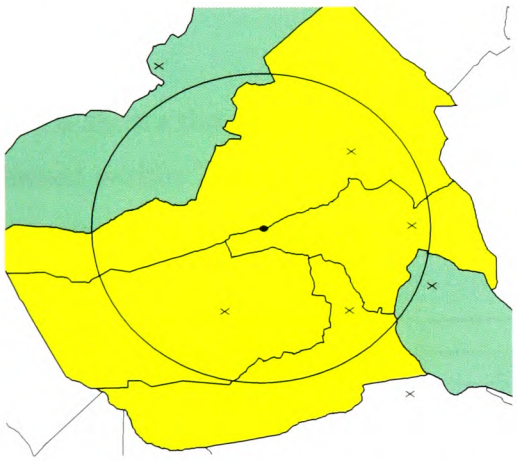


Figure 3.5(a): Centroid method

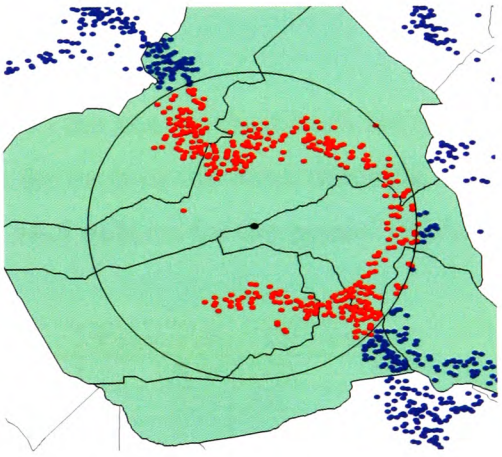


Figure 3.5(b) Postcode method

Figure 3.5: Comparison of centroid method and postcode method.

The diagram in figure 3.5(a) shows a hypothetical point source and the surrounding wards. Also included is a 2km buffer. Marked in each ward are the "population weighted" ward centroids (centroids have been allocated their respective central position considering the population distribution rather than the geographical centre of the ward i.e.

"geographical weighted" centroids). The centroid technique identifies those that are contained within the area of interest, in this case being the four wards highlighted in yellow within the 2km buffer. The postcode method examines all wards that are completely within the area at risk as well as those wards that cross the buffer boundary, thus an additional two wards are included (all six wards are highlighted green in figure 3.5(b)). Of the six wards selected, the total number of domestic delivery points for each postcode is noted for each ward. The corresponding total number of domestic delivery points within the buffer for each ward is noted (those postcodes that are highlighted red in figure 3.5(b) within 2km of the landfill site. Postcodes outside the area are highlighted in blue). The total number of domestic delivery points within the buffer for each ward is calculated as a percentage of the total number of delivery points in each ward and each proportion applied to the corresponding population figure of the ward and aggregated to obtain the total population at risk within 2km of the point source. This method assumes that the distribution of population in the exposed and non exposed areas within a ward is the same.

Table 3.6 shows the varying person years at risk (to the nearest thousand) for the analysis conducted within 2km of a landfill site in Wales for each of the three methods. The total person years at risk in Wales was approximately 56.5 million for the period 1982-2001.

	Wards	Enumeration Districts
Intersection Method	23,604,000	12,068,000
Centroid Method	8,836,000	9,271,000
Postcode Method	9,640,000	9,545,000

Table 3.6: Total person years at risk within 2km of landfill sites in Wales, 1982-2001.

As can be seen from table 3.6, the population at risk varies between the geographical unit used for each method. The difference in population at risk for the intersection method between using wards and EDs was 11.5 million person-years. The areas at risk were not entirely the same for each method so that different estimates were calculated for the population at risk for each method. The ward analysis using the intersection method

dilutes the estimate of the relative risk since it includes wards outside the area at risk (on a much greater scale than any of the other methods). The postcode method has the least change in population at risk between all methods. i.e. the difference between the population at risk for the postcode method for wards and enumeration districts is smallest between all three methods. Comparing the population at risk between the centroid method and postcode method, there were an additional 804,000 persons included in the ward analysis (approximately 40,000 per year) and an additional 274,000 persons included in the enumeration district analysis (approximately 14,000 per year). Clearly this is a large difference in the population at risk in the exposed area. Comparing the intersection method with the other methods, there is approximately 2.5 times the amount of population at risk in the exposed area which clearly identifies the large number of geographical units included in the analysis.

The numbers of observed cases within 2km of a landfill site were calculated, along with the numbers of cases that were expected in the area of risk, based on age-sex-deprivation age specific rates in Wales for the twenty year period 1982-2001. The standardised incidence ratio was calculated (the number of observed cases divided by the number of expected cases) and 95% confidence intervals examined for any significant increased or decreased risks. The lower and upper confidence limits were calculated using exact Poisson 95% confidence limits for the observed figures and then dividing by the number of expected cases.

It was postulated that the postcode method would provide the same results between any choice of geography unit due to the same area at risk being analysed whereas other methods would produce different results e.g. the centroid method (a geographical unit included in the analysis if its centroid was inside the buffer) will exclude particular wards but include enumeration districts in that ward if the centroid of the ward or ED was inside or outside the area at risk.

Other studies tend to use just one method and report conclusions based on this one method. The analysis presented here details the differences in population at risk and

resulting conclusions if other methods are used. Other studies do not tend to analyse the exact area of interest that they are examining due to using the centroid method whereby population at risk data are included in the exposed area when this is not the case. Also, the choice of geographical unit can cause different results and the way in which the observed cases are aggregated can have an effect of dampening any true effect, if one does exist. Also, postcode information is accurate to one metre resolution (average position of each postcode) by using the product CodePoint™ from Ordnance Survey for all current postcodes in operation in the UK (as of 2003) for this research whereas many past studies have accuracy to only 100 metres.

3.7. Results

Table 3.7 shows the analysis of cancers within 2km of landfill sites in Wales when using wards or EDs as the level of geography. Expected figures were calculated using five year age band, quintile of deprivation and sex specific rates for Wales as a whole, the usual methodology adopted by the WCISU. Both geography levels were used to determine if the choice of geographical unit affected the results.

The numbers in brackets in table 3.7 indicate the number of wards or enumeration districts (EDs) examined in the analysis and do not represent confidence intervals. The number of wards used for the postcode method and the intersection method are the same since both initially use all wards or EDs that intersect the 2km buffer. The second figure for the postcode method relates to the number of wards or EDs that contained a population at risk greater than zero – only a small part of a ward or ED may intersect the buffer and may include no population within this small area. As can be seen, the number of observed cases varies widely between each of the methods used.

Intersection method

This method produced the highest number of observed cases in the area at risk using both geographical units (3810 brain cancer cases compared with 1533 cases for the postcode

method using wards). There was a much larger population defined as being at risk using this method, hence the much higher number of observed cases in the analysis. There was also a large difference between the results when using wards compared with EDs (nearly twice as many cases were observed when using wards compared with EDs). No results were significant when using this method.

Centroids method

There were a lower number of wards and EDs examined in the analysis compared to the previous method since the ward or ED was not included in the analysis if its centroid was outside the 2km buffer. Comparing the ward and ED level analysis, there was a difference in the number of observed cases in the analysis, although not as large a difference as with the intersection method. A conflicting increased risk and decreased risk (or vice versa) were found for some cancers when using wards instead of enumeration districts. For example, leukaemia shows a 1% significant increased risk when analysing at ward level but displays a 3% decreased risk when analysing at ED level. However, both results are non-significant. The geography level used has clearly made a difference to the observed cases, expected cases and SIRs obtained in the analysis.

Postcode method

Table 3.7 shows no significant increased risks within 2km of a landfill site in Wales for the twenty year period 1982-2001. The majority of Standardised Incidence Ratios (SIR) were close to 1. The large numbers of observed cases within 2km of the landfill sites in Wales ensures small confidence intervals and overcomes the problem of an increase of one or two cases in the exposed area dramatically affecting the overall result. Note that for the postcode method, very similar results were obtained irrespective of the choice of geography level as was expected.

INTERSECTION		WARDS (276)				EDs (1380)			
Cancer	Sex	Obs	Exp	SIR	95% CI	Obs	Exp	SIR	95% CI
Brain Cancer	Males	1881	1859.60	1.01	(0.97, 1.06)	1042	1064.57	0.98	(0.92, 1.04)
	Females	1929	1905.31	1.01	(0.97, 1.06)	1107	1081.18	1.02	(0.96, 1.09)
	Persons	3810	3764.90	1.01	(0.98, 1.04)	2149	2145.75	1.00	(0.96, 1.04)
Leukaemia	Males	1715	1688.54	1.02	(0.97, 1.06)	972	966.50	1.01	(0.94, 1.07)
	Females	1304	1319.61	0.99	(0.94, 1.04)	716	749.30	0.96	(0.89, 1.03)
	Persons	3019	3008.15	1.00	(0.97, 1.04)	1688	1715.80	0.98	(0.94, 1.03)
CENTROIDS		WARDS (100)				EDs (937)			
Cancer	Sex	Obs	Exp	SIR	95% CI	Obs	Exp	SIR	95% CI
Brain Cancer	Males	695	706.92	0.98	(0.91, 1.06)	740	743.12	1.00	(0.93, 1.07)
	Females	721	729.92	0.99	(0.92, 1.06)	763	758.44	1.01	(0.94, 1.08)
	Persons	1416	1436.83	0.99	(0.93, 1.04)	1503	1501.56	1.00	(0.95, 1.05)
Leukaemia	Males	667	653.97	1.02	(0.94, 1.10)	678	679.04	1.00	(0.92, 1.08)
	Females	517	516.17	1.00	(0.92, 1.09)	497	526.83	0.94	(0.86, 1.03)
	Persons	1184	1170.14	1.01	(0.96, 1.07)	1175	1205.87	0.97	(0.92, 1.03)
POSTCODE METHOD		WARDS (276, 239)				EDs (1380,1231)			
Cancer	Sex	Obs	Exp	SIR	95% CI	Obs	Exp	SIR	95% CI
Brain Cancer	Males	733	760.20	0.96	(0.90, 1.04)	733	763.18	0.96	(0.89, 1.03)
	Females	800	784.23	1.02	(0.95, 1.09)	800	780.17	1.03	(0.96, 1.10)
	Persons	1533	1544.44	0.99	(0.94, 1.04)	1533	1543.35	0.99	(0.94, 1.04)
Leukaemia	Males	708	696.72	1.02	(0.94, 1.09)	708	696.21	1.02	(0.94, 1.09)
	Females	525	548.22	0.96	(0.88, 1.04)	525	542.36	0.97	(0.89, 1.05)
	Persons	1233	1244.93	0.99	(0.94, 1.05)	1233	1238.56	1.00	(0.94, 1.05)

Table 3.7: Analysis of cancers within 2km of a landfill site in Wales for all methods.

Comparing the postcode method results with previous studies, brain cancer results were very similar to the study by Jarup et al (2002). Adjusted rate ratios for brain cancer in the study by Jarup resulted in non-significant rate ratios of 0.99 (within 2km of all landfill sites and all special waste landfill sites that operated at any time during the study period). This study provided non-significant SIRs of 0.99 using ward and ED analysis for brain cancer. For adult leukaemia (greater than 14 years of age) in Jarup's study, rate ratios of 0.99 were calculated whereas in this study, SIRs of 0.99 (ward and ED analysis) were calculated if using only those aged over 14 years, again very similar to Jarup's results.

In summary, it can be seen that the choice of geography level and method affects the population at risk in the analysis and the resulting number of observed and expected cases. However, the results here show small changes in SIRs, probably due to there

being no evidence for a change in risk in these areas. Single site analysis is explored in section 3.8 to examine the effect that this has on the method used for analysis since at WCISU only one landfill site may be the subject of concern. The exposed areas that were analysed for each of the geography levels for the intersection and centroid method were clearly different. There was a large difference in the number of observed and expected cases for the cancer sites examined and SIRs vary depending on the choice of geography level. It can be seen that the postcode method gives very similar results irrespective of the choice of geography level. However, the analysis presented so far did not depend on whether the landfill site was in operation at the time that the case was diagnosed. This is explored further in section 3.10. The primary focus in this theme has been on methodological aspects rather than epidemiological factors.

Extrapolation projects the 1991 Census population figures backward to 1982 and forwards to 2001. The sex and five year age band proportional change from each year to 1991 was noted at Local Health Board level and these proportions applied to the corresponding 1991 Census population figures at ward or ED level in their respective local health board to all years from 1982 to 1990 and 1992 to 2001. The mid-year Local Health Board population estimates are available from ONS. However, extrapolation did not affect any of the results. Table 3.8 shows the comparison of expected numbers within 2km of landfill sites in Wales when using extrapolation of population using the postcode method and when not using extrapolation.

POSTCODE METHOD			
Cancer	**Sex**	**No extrapolation**	**Extrapolation**
Brain Cancer	**Males**	763.18	766.53
	Females	780.17	783.58
	Persons	1543.35	1550.10
Leukaemia	**Males**	696.21	699.43
	Females	542.36	544.90
	Persons	1238.56	1244.33

Table 3.8: Comparison of expected numbers using extrapolation of population.

Table 3.8 does not show the resulting SIR and 95% confidence interval since they generally remained the same (just two of the six results changed at 2 decimal places) to those obtained in table 3.7. The number of expected cases changed only slightly, thus the very small change in the corresponding 95% CIs. However, if there was a large population increase in a particular ED due to a new housing development for example, then the population in that particular ED would increase dramatically, however the method shown here assumes an equal increase or decrease in that ED from its corresponding local health board change from year to year. To summarise, taking extrapolation into account did not appear to influence the results using the methodologies employed in this study.

3.8. Analysis of an increased risk around one landfill site

When a cluster enquiry is received at WCISU, only one landfill site is generally of interest. Consider a single unidentified landfill site in Wales referred to as landfill site *X*. The three methods were then compared using this area within 2km of the landfill site.

The actual population figures in Wales were used in the analysis along with the observed numbers of cancer cases in Wales to obtain the sex-age-deprivation specific rates to calculate the expected numbers within 2km of landfill *X*. Table 3.9 shows the results using the three methods:

	Intersection Method				
	Obs	**Exp**	**Population-yrs**	**SIR**	**95% CI**
Males	15	11.13	62000	1.35	(0.755, 2.223)
Females	13	8.50	69000	1.53	(0.814, 2.614)
Persons	28	19.63	131000	1.43	(0.948, 2.062)
	Centroid Method				
	Obs	**Exp**	**Population-yrs**	**SIR**	**95% CI**
Males	7	5.99	32000	1.17	(0.469, 2.406)
Females	11	4.72	37000	2.33	(1.163, 4.167)
Persons	18	10.72	69000	1.68	(0.995, 2.654)
	Postcode Method				
	Obs	**Exp**	**Population-yrs**	**SIR**	**95% CI**
Males	10	7.61	41000	1.31	(0.630, 2.417)
Females	12	5.96	46000	2.01	(1.041, 3.519)
Persons	22	13.57	87000	1.62	(1.016, 2.455)

Table 3.9: Analysis of the three methods within 2km of landfill X.

As can be seen in table 3.9, the three methods show an increased risk within 2km of the landfill site (43% for the intersection method, 68% for the centroid method and 62% for the postcode method). However, the centroid method shows a borderline significant result, the intersection method shows a non-significant result and the postcode method shows a significant result. There is also a large difference in the population-years used in the analysis. The centroid method can distort a real significant increased risk due to the inclusion or exclusion of specific geographical units. The centroid method has used population and observed cases outside the buffer of 2km due to the position of the centroid being within the buffer. Also, cases and population at risk within the buffer are not included due to the centroid of the ED being outside the buffer. The centroid method and intersection method (in particular) have "dampened" the effect of the increased risk within 2km of the landfill site. The intersection method also includes a larger area outside the buffer that was included in the analysis. This example clearly shows the importance of analysing the exact area of interest and consequently the actual population "at risk". The population at risk used in the analysis varies between each method. The population at risk using the intersection method is nearly twice that compared with the centroid method. The difference between the centroid method and postcode method is 18,000 person years – nearly 1000 per year. Even though this figure is a small difference per year, the resulting SIR and CIs show the effect that this has on the results.

3.9. Determining the "unexposed" population – Comparison of results

The results in the previous section used the population of all Wales to calculate expected figures in the exposed areas of interest. i.e. comparing rates in the whole of Wales with the rates in the exposed area. This is standard practice at the WCISU and many other cancer registries in the UK. However, in the majority of cases, only one landfill site is usually analysed at the WCISU. The following analysis compares this "standard practice" with that of calculating expected numbers in the "exposed" region using only the "unexposed" population at risk (as opposed to all Wales) when using the postcode method. The geographical units used were enumeration districts due to the similar figures that were observed previously for both geographical units when using the

postcode method. Sex-age-deprivation specific rates of the "unexposed" population were used to calculate the expected numbers within 2km of a landfill site. The first "unexposed" population was based on all Wales as previously used. The second "unexposed" population was based on those not living within 2km of a major landfill site in Wales. Table 3.10 summarises the results.

			Unexposed: WALES			Unexposed: >2km		
			Int	Cent	Pcode	Int	Cent	Pcode
MALES	Brain Cancer	Observed	1042	740	733	1042	740	733
		Expected	1064.57	743.12	763.18	1073.06	744.99	771.08
		SIR	0.98	1.00	0.96	0.97	0.99	0.95
	Leukaemia	Observed	972	678	708	972	678	708
		Expected	966.50	679.04	696.21	964.62	678.96	694.44
		SIR	1.01	1.00	1.02	1.01	1.00	1.02
FEMALES	Brain Cancer	Observed	1107	763	800	1107	763	800
		Expected	1081.18	758.44	780.17	1076.89	759.92	778.31
		SIR	1.02	1.01	1.03	1.03	1.00	1.03
	Leukaemia	Observed	716	497	525	716	497	525
		Expected	749.30	526.83	542.36	757.58	533.24	546.79
		SIR	0.96	0.94	0.97	0.95	0.93	0.96
PERSONS	Brain Cancer	Observed	2149	1503	1533	2149	1503	1533
		Expected	2145.75	1501.56	1543.35	2149.95	1504.91	1549.39
		SIR	1.00	1.00	0.99	1.00	1.00	0.99
	Leukaemia	Observed	1688	1175	1233	1688	1175	1233
		Expected	1715.80	1205.87	1238.56	1722.20	1212.20	1241.23
		SIR	0.98	0.97	1.00	0.98	0.97	0.99

Table 3.10: Comparison of results using different "unexposed" populations.

In summary, table 3.10 shows that there was very little difference in the results when using both methods for determining those not exposed. The unexposed population changed from 100% (all Wales) to approximately 83% for those living greater than 2km from a landfill site. No results were significant. This was probably due to there being no apparent increased risk in the exposed area of interest in the results and the population figure at risk in the "exposed" areas being very small compared to those not exposed. However, as a general principle, results should be based on those truly not exposed. Thus a recommendation to the WCISU is that all Wales analysis should not be used for future work when calculating expected numbers. In fact, perhaps a distance greater than 5km from landfill sites should have been used as the population in the unexposed area to

calculate the expected figures due to the ambiguous evidence for the risk being restricted to 2km within landfill sites. Although similar results were obtained, the unexposed population at risk should not contain the exposed population at risk since if a true increased risk exists around landfill sites then using all Wales rates will dilute the effect. However, comparing both tables, like for like, then SIRs to 2 decimal places are very similar. In fact, of the 18 comparisons of SIRs that can be made, 16 have a difference of less than 0.01 and 2 have a difference of between 0.01 and 0.02. i.e. no apparent effect is shown here.

3.10. Operation dates of landfill sites

Table 3.11 extends the analysis for landfill sites using the postcode method (enumeration districts) further so that the population at risk only includes those people who were "exposed" during the time of operation of the landfill sites in Wales and the unexposed population were those that did not live within 2km of a landfill site or were never exposed to the landfill site if they did live within 2km of a landfill site in Wales.

Cancer	Sex	Obs	Exp	SIR	95% CI
Brain Cancer	Males	404	433.54	0.93	(0.84, 1.03)
	Females	438	441.49	0.99	(0.90, 1.09)
	Persons	842	875.03	0.96	(0.90, 1.03)
Leukaemia	Males	404	389.14	1.04	(0.94, 1.14)
	Females	299	304.98	0.98	(0.87, 1.10)
	Persons	703	694.12	1.01	(0.94, 1.09)

Table 3.11: Results of analysis when taking into account the operation times of the landfill sites.

Comparing the results in table 3.11 with the corresponding results in table 3.10, the SIRs for leukaemia have increased but are still not significant and the SIRs of brain cancer have decreased. The number of observed cases have fallen by nearly a half for the cancer sites examined compared with the previous analysis. These results compare favourably with studies quoted earlier.

3.11. Latency of cancers

The time between first exposure to a cancer-causing agent and diagnosis of the disease is called the latency (or latent) period. Cancer in general is thought to have a latency period of between 15 and 20 years (Southern Medical Services Ltd, 2004). However this varies between cancer sites.

Salvati et al (2003) claim that the mean latent period for brain cancer is approximately 12 years. Other studies have shown similar latency periods with a range between 1 and 26 years. The Radiation Protection Board state that the latent period for leukaemia is between 2 and 10 years. The large variation in years is due to the exposure dose. For example, a person exposed to a single large dose of ionizing radiation will generally result in a short latent period from exposure to diagnosis of cancer whereas a person exposed to a low dose of ionizing radiation will tend to have longer latent periods from exposure to diagnosis. The latency periods noted here will be used in future analysis.

The hypothesis is of an increased risk within 2km of a particular landfill site (exposed cases) compared with the risk more than 2km away from the landfill site (unexposed cases).

To take into account latency, figure 3.6 shows the location of exposed leukaemia cases (green) within 2km of a hypothetical landfill site and unexposed leukaemia cases (blue) outside this area for the diagnosis period 1982-2001. Assume that the landfill site was opened in 1991 but closed in 1996. The exposed cases in figure 3.6 have been allocated their diagnosis years. All cases not within 2km are contained in the calculation for the number of expected cases within 2km of the landfill site. The dates in brackets indicate the period that the cases were exposed to the landfill site. i.e. between 2 and 10 years from diagnosis. The exposed cases that are marked with squares indicate the "true" exposed cases since they were exposed to the radiation during the time that the landfill site was in operation. The remaining cases within 2km could not have been exposed to the radiation since their "exposure" period preceded the opening date of the landfill site. Hence, only the "true" exposed cases are to be included in the analysis within 2km of the

landfill site. The population at risk in the area is adapted to account for the latency period. i.e. cases diagnosed between 1982 and 1993 (and corresponding populations at risk) and those that reside within 2km of the landfill site are not to be included in the analysis since their "exposure" period was before the landfill site opened. Hence only the cases diagnosed between 1993 and 2001 are included in the "exposure" period.

Age must also be taken into account since a 4 year old child diagnosed in 2001 would have an "exposure" period between 1997 and 1999 (as the child would not have been born before 1997). Thus this case is also not in the "exposure" period assuming that no one was exposed to radiation when the landfill site closed.

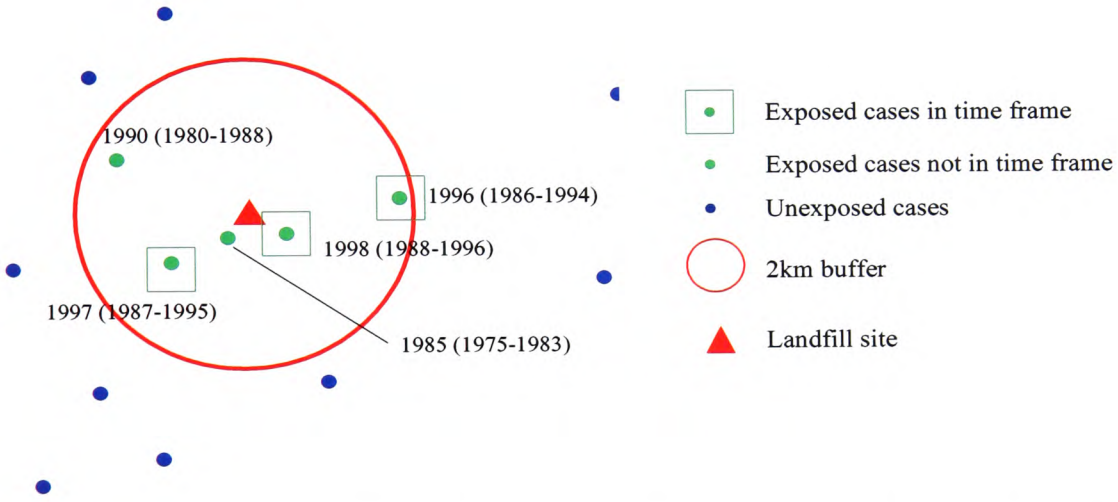


Figure 3.6: Assigning exposed cases within a landfill site (operational between 1991 and 1996) using latency periods.

Table 3.12 shows the resultant exposure matrix (shaded grey) for the leukaemia cases of a certain age at diagnosis and year of diagnosis. The cells not exposed (ne) to the radiation are therefore not included in the analysis living within 2km of a landfill site. The dates in the matrix correspond to the years of exposure for a case diagnosed in a particular year at a particular age (in years).

		Age in years											
		0	1	2	3	4	5	6	7	8	9	10	>10
Year of diagnosis	1982	ne	ne	ne	ne	ne	ne	ne	ne	ne	ne	ne	ne
	1983	ne	ne	ne	ne	ne	ne	ne	ne	ne	ne	ne	ne
	1984	ne	ne	ne	ne	ne	ne	ne	ne	ne	ne	ne	ne
	1985	ne	ne	ne	ne	ne	ne	ne	ne	ne	ne	ne	ne
	1986	ne	ne	ne	ne	ne	ne	ne	ne	ne	ne	ne	ne
	1987	ne	ne	ne	ne	ne	ne	ne	ne	ne	ne	ne	ne
	1988	ne	ne	ne	ne	ne	ne	ne	ne	ne	ne	ne	ne
	1989	ne	ne	ne	ne	ne	ne	ne	ne	ne	ne	ne	ne
	1990	ne	ne	ne	ne	ne	ne	ne	ne	ne	ne	ne	ne
	1991	ne	ne	ne	ne	ne	ne	ne	ne	ne	ne	ne	ne
	1992	ne	ne	ne	ne	ne	ne	ne	ne	ne	ne	ne	ne
	1993	ne	ne	91	91	91	91	91	91	91	91	91	91
	1994	ne	ne	92	91-92	91-92	91-92	91-92	91-92	91-92	91-92	91-92	91-92
	1995	ne	ne	93	92-93	92-93	91-93	91-93	91-93	91-93	91-93	91-93	91-93
	1996	ne	ne	94	93-94	92-94	91-94	91-94	91-94	91-94	91-94	91-94	91-94
	1997	ne	ne	95	94-95	93-95	92-95	91-95	91-95	91-95	91-95	91-95	91-95
	1998	ne	ne	96	95-96	94-96	93-96	92-96	91-96	91-96	91-96	91-96	91-96
	1999	ne	ne	ne	96	95-96	94-96	93-96	92-96	91-96	91-96	91-96	91-96
	2000	ne	ne	ne	ne	96	95-96	94-96	93-96	92-96	91-96	91-96	91-96
	2001	ne	ne	ne	ne	ne	96	95-96	94-96	93-96	92-96	91-96	91-96

Table 3.12: Exposure matrix for age and year at diagnosis for cases within 2km of a landfill site.

Only the cases highlighted in grey are included in the analysis along with the corresponding population at risk. The population that were within the buffer but not at risk following the latency period, "ne", are now included in the unexposed population to calculate expected figures (if using unexposed analysis as opposed to all Wales analysis). Population data by enumeration district is only available in five-year age bands for a particular Census year (1991 Census in this case) and different years and ages are included in the exposure matrix in table 3.12. Thus, the proportion of the population in Wales for each single year of age was applied to the corresponding five year age band by ward to obtain population figures by single year of age by ward. The postcode method was used to identify the exact population at risk within the area of interest.

To examine the effect that the latency period had on the population at risk and resultant observed and expected cases, the focused space-time scan statistic was used to identify a cluster around a landfill site in Wales for a particular time period between 1992 and 2001

for brain cancer and leukaemia. These two clusters were examined in terms of resulting SIRs using various techniques described earlier and taking the latency period into account. SIRs were calculated along with 95% CIs and p-values. In total, four methods were used to analyse the datasets; centroid method from the spatial scan statistic results, the postcode method (all Wales analysis) and the postcode method following the latency period effect (all Wales analysis and unexposed population analysis to calculate expected figures). Note that it would have been ideal to use the spatial scan statistic using postcode data but no age-sex population breakdown was available by postcode.

Brain cancer

A focused space-time statistic was used (the clustering algorithm that was selected as the most appropriate local clustering algorithm as judged from the first theme) to locate clusters around any of the landfill sites in Wales for the period 1992-2001. A most likely cluster was located around a landfill site in Cardiff for the period 1992-1994. Table 3.13 shows those cases truly exposed during the diagnosis period 1992-1994 within 4.8km of the landfill site and table 3.14 shows the results of the analysis from the scan statistic (centroids) using the postcode method. The landfill site was in operation between 1980 and 1989.

Age (years)	1992	1993	1994
0	ne	ne	ne
1	ne	ne	ne
2	ne	ne	ne
3	89	ne	ne
4	88-89	89	ne
5	87-89	88-89	89
6	86-89	87-89	88-89
7	85-89	86-89	87-89
8	84-89	85-89	86-89
9	83-89	84-89	85-89
10	82-89	83-89	84-89
11	81-89	82-89	83-89
12	81-89	81-89	82-89
13	81-89	81-89	81-89
>13	81-89	81-89	81-89

Table 3.13: Exposure matrix for those diagnosed between 1992 and 1994 within 4.8km of the landfill site.

A significant increased risk of 66% was obtained using the method of population weighted centroids. This fell by 1% when using the postcode method. No cases within 4.8km of the landfill site were not exposed (i.e. no cases of brain cancer were allocated into the cells in table 3.13 with "ne"), hence the same number of observed cases. However, since the population at risk has decreased by those cells labelled "ne", the increased risk has risen further to 67%. This rises even further to 72% when comparing the exposed cases to the unexposed cases.

Method	Sex	Observed	Expected	SIR	95% Conf. Interval
Centroids	Male	52	35.05	1.48	(1.11, 1.95)
	Female	73	40.06	1.82	(1.43, 2.29)
	Persons	125	75.11	1.66	(1.39, 1.98)
Postcode method (Unexposed - Wales)	Male	49	34.00	1.44	(1.07, 1.91)
	Female	71	38.81	1.83	(1.43, 2.31)
	Persons	120	72.82	1.65	(1.37, 1.97)
Postcode method (Unexposed - Wales) & LP	Male	49	33.63	1.46	(1.08, 1.93)
	Female	71	38.37	1.85	(1.45, 2.33)
	Persons	120	72.00	1.67	(1.38, 1.99)
Postcode method (Unexposed >4.8km) & LP	Male	49	32.90	1.49	(1.10, 1.97)
	Female	71	36.74	1.93	(1.51, 2.44)
	Persons	120	69.63	1.72	(1.43, 2.06)

LP: Latency period

Table 3.14: Analysis within 4.8km of Bute Dock landfill site, Cardiff, 1992-1994.

Leukaemia

Table 3.15 shows the exposure matrix for those cases truly exposed within 42.4km of a landfill site located in Pembrokeshire between 1996 and 1998 (the diagnosis period of the cluster). The site was in operation between 1961 and 1986. The latency period used was between 2 and 10 years. Note that the geographical size of the cluster is very large at over 42km and hence the question as to whether this should actually be termed a "cluster". For illustration purposes, the analysis is presented here. The problem here is that the actual observed cluster was 10 years after the site closed and hence only those cases diagnosed in 1996 and aged over 10 years were truly exposed.

Age (years)	1996	1997	1998
0	ne	ne	ne
1	ne	ne	ne
2	ne	ne	ne
3	ne	ne	ne
4	ne	ne	ne
5	ne	ne	ne
6	ne	ne	ne
7	ne	ne	ne
8	ne	ne	ne
9	ne	ne	ne
>9	1986	ne	ne

Table 3.15: Exposure matrix for leukaemia between 1996 and 1998 within 42.4km of the landfill site.

Table 3.16 compares the methods when analysing leukaemia.

Method	Sex	Observed	Expected	SIR	95% Conf. Interval
Centroids	Male	54	25.11	2.15	(1.62, 2.81)
	Female	36	18.62	1.93	(1.35, 2.68)
	Persons	90	43.73	2.06	(1.66, 2.53)
Postcode method	Male	46	24.78	1.86	(1.36, 2.48)
	Female	38	18.40	2.07	(1.46, 2.84)
	Persons	84	43.18	1.95	(1.55, 2.41)
Postcode method (Unexposed - Wales)	Male	20	7.92	2.52	(1.54, 3.90)
	Female	8	5.84	1.37	(0.59, 2.70)
	Persons	28	13.76	2.03	(1.35, 2.94)
Postcode method (Unexposed > 42.4km)	Male	20	8.22	2.43	(1.49, 3.76)
	Female	8	6.06	1.32	(0.57, 2.60)
	Persons	28	14.28	1.96	(1.30, 2.83)

Table 3.16: Analysis within 42.4km of the landfill site, 1996-1998.

Using the population weighted centroids method, there was a significantly increased risk of 106% based on 90 cases which decreased to a significantly increased risk of 95% using the postcode method based on 84 cases. When taking the latency period into account the numbers of observed cases decreased dramatically to just 28 cases. The significantly increased risks were still evident although the confidence intervals were wider compared with not taking the latency period into account.

3.12. Electric power line analysis

The previous analysis was explored further using a linear source, i.e. electric power lines in Wales as opposed to point sources. Past studies tend to use various distances from electric power lines to determine risk ranging from 100 metres to 1000 metres. A distance of 100 metres was primarily used to investigate the risk of various cancers in this area of research. However, further work explores the definition of risk ranging from 100 metres up to 1000 metres.

The three methods studied in the previous section were used to determine whether an increased risk existed within 100 metres of an electric power line in Wales. Table 3.17 shows the populations at risk (all ages, to the nearest thousand) for the electric power line analysis for all three methods.

	Wards	Enumeration Districts
Intersection Method	29,770,000	9,815,000
Centroids Method	820,000	940,000
Postcode Method	842,000	812,000

Table 3.17: Person years at risk within 100 metres of an electric power line in Wales, 1982-2001.

The intersection method had over ten times the population at risk compared with the other two methods when using EDs as the geographical unit. Table 3.18 shows the results of this analysis for the postcode method (Pcode), intersection method (Int) and Centroid method (Cent) using EDs as the geographical unit. The unexposed population at risk was defined as those not exposed to the electric power lines i.e. those living greater than 100 metres from an electric power line (over 132kV) in Wales. Table 3.19 compares the number of same patients (observed cases) included in the exposed area for the three population estimation techniques.

			Unexposed: WALES			Unexposed: >100 metres		
			Int	Cent	Pcode	Int	Cent	Pcode
MALES	Leukaemia	Observed	704	58	46	704	58	46
		Expected	700.03	61.86	53.00	700.59	61.86	53.07
		SIR	1.01	0.94	0.87	1.00	0.94	0.87
	Childhood Cancer	Observed	128	17	14	128	17	14
		Expected	133.98	13.57	11.86	135.87	13.52	11.80
		SIR	0.96	1.25	1.18	0.94	1.26	1.19
FEMALES	Leukaemia	Observed	531	53	38	531	53	38
		Expected	523.28	47.50	39.29	521.99	47.35	39.29
		SIR	1.01	1.12	0.97	1.02	1.12	0.97
	Childhood Cancer	Observed	97	3	4	97	3	4
		Expected	113.97	11.31	9.89	117.16	11.47	10.01
		SIR	0.85	0.27	0.40	0.83	0.26	0.40
PERSONS	Leukaemia	Observed	1235	111	84	1235	111	84
		Expected	1223.31	109.35	92.29	1222.58	109.20	92.36
		SIR	1.01	1.02	0.91	1.01	1.02	0.91
	Childhood Cancer	Observed	225	20	18	225	20	18
		Expected	247.96	24.88	21.74	253.03	24.99	21.81
		SIR	0.91	0.80	0.83	0.89	0.80	0.83

Table 3.18: Comparison of results using different "unexposed" populations.

Childhood Cancer			
	Intersection	**Centroid**	**Postcode**
Intersection	-	18 of 225	18 of 225
Centroid	18 of 20	-	4 of 20
Postcode	18 of 18	4 of 18	-
Leukaemia			
	Intersection	**Centroid**	**Postcode**
Intersection	-	102 of 1235	84 of 1235
Centroid	102 of 111	-	22 of 111
Postcode	84 of 84	22 of 84	-

Table 3.19: Number of cases included in the exposed region using various methods.

Table 3.18 shows two significantly decreased risks (highlighted yellow), for female childhood cancer when using the centroid method. The significant results are not present for the intersection method or postcode method. Clearly, it can be seen from table 3.18 that the intersection method produced a large difference in the number of observed cases

compared with the postcode method and centroid method. This was due to the "exposed" area surrounding the electric power lines being very small, at 100m, and the whole ward or ED was included in the analysis if it intersected with an electric power line. However, what is not shown here is the difference in the cases selected using the 'centroid' method. Many cases included in the intersection analysis were over 100 metres away and many cases within 100 metres were not included in the 'centroid' analysis. If the actual cases in the analysis for each of the methods were compared then the postcode method included just 22 cases that were also included in the exposed region using the centroid method (from a total of 84 cases). This is shown in table 3.19. The electric power line analysis results (expected numbers and SIRs) were very similar between the choice of the unexposed population at risk due to the small population figure within 100 metres (only 1.4% of the population resided within 100 metres of an electric power line in Wales).

For electric power line analysis, there were few cases that lived within 100 metres of an electric power line, especially for childhood cancer. To enable robust estimates this section examines the SIR within electric power lines in steps of 100 metres up to 1000 metres. Table 3.20 shows the total population at risk (to the nearest thousand) that were exposed within the specified distances for the period 1982-2001.

Distance	Adults	Children, 0-14 years
<=100m	812,000	167,000
<=200m	2,184,000	448,000
<=300m	3,641,000	748,000
<=400m	5,491,000	1,129,000
<=500m	7,422,000	1,523,000
<=600m	9,470,000	1,943,000
<=700m	11,467,000	2,346,000
<=800m	13,590,000	2,774,000
<=900m	15,702,000	3,204,000
<=1000m	17,652,000	3,597,000
>1000m	38,850,000	7,246,000

Table 3.20: Total person years at risk within distances for power line analysis, 1982-2001.

All results compare the "exposed" population at risk to those not living within 1000 metres of an electric power lines in Wales using age-sex-deprivation specific rates to calculate expected numbers. Table 3.21 shows the results of this analysis.

There was a significant decreased risk of female childhood cancer within 400 metres of a power line and a significant decreased risk of female childhood cancer within 1000 metres of a power line. Note that the upper confidence limit for female childhood cancer is just over unity for the other distances. There were significant increased risks for male leukaemia within 400 metres of electric power lines in Wales which continued up to within 1000 metres of a power line. Females also showed a significant increased risk within 800 metres, 900 metres and 1000 metres of a power line in Wales but the risk was actually highest at approximately 400 metres from the electric power lines for males and females combined.

	Males				Females				Persons			
Leukaemia	**Obs**	**Exp**	**SIR**	***95% Conf Int***	**Obs**	**Exp**	**SIR**	***95% Conf Int***	**Obs**	**Exp**	**SIR**	***95% Conf Int***
<=100m	46	51.75	0.89	(0.65, 1.19)	38	38.57	0.99	(0.70, 1.35)	84	90.33	0.93	(0.74, 1.15)
<=200m	154	142.20	1.08	(0.92, 1.27)	105	107.44	0.98	(0.80, 1.18)	259	249.64	1.04	(0.91, 1.17)
<=300m	268	238.95	1.12	(0.99, 1.26)	183	180.65	1.01	(0.87, 1.17)	451	419.60	1.07	(0.98, 1.18)
<=400m	412	363.59	1.13	(1.03, 1.25)	309	277.75	1.11	(0.99, 1.24)	721	641.34	1.12	(1.04, 1.21)
<=500m	547	494.52	1.11	(1.02, 1.20)	419	381.70	1.10	(1.00, 1.21)	966	876.21	1.10	(1.03, 1.17)
<=600m	695	634.23	1.10	(1.02, 1.18)	535	491.22	1.09	(1.00, 1.19)	1230	1125.46	1.09	(1.03, 1.16)
<=700m	856	771.74	1.11	(1.04, 1.19)	643	599.41	1.07	(0.99, 1.16)	1499	1371.15	1.09	(1.04, 1.15)
<=800m	1000	919.08	1.09	(1.02, 1.16)	773	713.01	1.08	(1.01, 1.16)	1773	1632.09	1.09	(1.04, 1.14)
<=900m	1171	1065.33	1.10	(1.04, 1.16)	890	827.44	1.08	(1.01, 1.15)	2061	1892.77	1.09	(1.04, 1.14)
<=1000m	1316	1201.70	1.10	(1.04, 1.16)	998	934.72	1.07	(1.00, 1.14)	2314	2136.42	1.08	(1.04, 1.13)
Childhood	**Obs**	**Exp**	**SIR**	***95% Conf Int***	**Obs**	**Exp**	**SIR**	***95% Conf Int***	**Obs**	**Exp**	**SIR**	***95% Conf Int***
<=100m	14	12.34	1.13	(0.62, 1.90)	4	10.47	0.38	(0.10, 0.98)	18	22.81	0.79	(0.47, 1.25)
<=200m	32	32.06	1.00	(0.68, 1.41)	15	27.69	0.54	(0.30, 0.89)	47	59.75	0.79	(0.58, 1.05)
<=300m	53	53.17	1.00	(0.75, 1.30)	28	46.19	0.61	(0.40, 0.88)	81	99.36	0.82	(0.65, 1.01)
<=400m	83	79.32	1.05	(0.83, 1.30)	49	69.66	0.70	(0.52, 0.93)	132	148.99	0.89	(0.74, 1.05)
<=500m	102	106.34	0.96	(0.78, 1.16)	78	93.89	0.83	(0.66, 1.04)	180	200.23	0.90	(0.77, 1.04)
<=600m	132	135.11	0.98	(0.82, 1.16)	101	119.65	0.84	(0.69, 1.03)	233	254.76	0.91	(0.80, 1.04)
<=700m	152	163.14	0.93	(0.79, 1.09)	129	144.49	0.89	(0.75, 1.06)	281	307.63	0.91	(0.81, 1.03)
<=800m	177	193.08	0.92	(0.79, 1.06)	157	171.03	0.92	(0.78, 1.07)	334	364.11	0.92	(0.82, 1.02)
<=900m	211	222.96	0.95	(0.82, 1.08)	170	197.50	0.86	(0.74, 1.00)	381	420.46	0.91	(0.82, 1.00)
<=1000m	225	250.52	0.90	(0.78, 1.02)	189	221.31	0.85	(0.74, 0.98)	414	471.83	0.88	(0.79, 0.97)

Table 3.21: Electric power line analysis (Exposed v Unexposed >1000m).

The study by Draper et al (2005) found significant increased risks for childhood leukaemia cases living within 200 metres (relative risk = 1.69) and between 200 metres and 600 metres (relative risk = 1.23) of electric power lines in the UK compared with those living greater than 600 metres from electric power lines. It was interesting to note that there were significant increased risks of male leukaemia within 400 metres and significant increased risks of female leukaemia within 800 metres of electric power lines but significant decreased risks of female childhood cancer for distances within 400 metres of electric power lines in Wales. The highest of the risks were for male and female leukaemia cases living within 400 metres of an electric power line in Wales. To enable a direct comparison with Draper et al., the dataset for leukaemia was cut to those aged between 0 and 14 for cases within 400 metres of electric power lines in Wales and compared with those living greater than 600 metres from electric power lines. However, no significant results were obtained (Males 95% CI (0.84, 1.78) based on 30 cases, Females 95% CI (0.33, 1.06) based on 13 cases). In the UKCSSI study (2000), a non significant odds ratio of 0.92 was calculated for all childhood cancers within 50 metres of a power line in the UK. This study provided a SIR of 0.83 for all childhood cancers (persons) within 100 metres of a power line in Wales (females were significantly lower) compared with those living greater than 1000 metres from electric power lines. Note the slight difference in results for those living within 100 metres in table 3.21 compared with table 3.18 due to the unexposed populations being greater than 1000 metres and greater than 100 metres respectively.

For adult cancers, a border line significant result was found for all haematological malignancies within 50 metres of a power line in the study by Youngson et al (1991) in Yorkshire, UK (odds ratio = 1.29). This study produced non-significant SIRs for leukaemia within 100 metres of a power line in Wales.

In general, results are not directly comparable to other studies since all studies tend to use differing distances for levels of exposure. Thus there is difficulty in determining whether an increased risk does actually exist near power lines.

Note that other factors mentioned previously have not been taken into account which could affect the results. e.g. unmeasured confounders, quality of diagnosis of disease, classification of disease, accuracy of population estimates, variations due to chance or residential location only taken into account. Again, the focus has been on the implications of the methodology on the overall findings and to provide contextual information for more detailed epidemiological investigations.

3.13. Conclusions

To conclude, the postcode method appeared to produce similar results irrespective of the geographical unit that was used for this analysis. The difference in SIR obtained was between 0.00 and 0.01 for the cancer sites examined within 2km of landfill sites in Wales. The difference in SIR obtained using the intersection method varied between 0.00 and 0.03 and between 0.00 and 0.06 for the centroid method for the analysis presented here. Extrapolation of population did not alter the resulting conclusions when using the postcode method and EDs. Extrapolation generally slightly "dampened" the results i.e. moved the SIR closer to unity. The 95% CI stayed exactly the same (to 2 decimal places) for the majority of results. The choice of what was used to define the unexposed population for calculation of the number of expected cases (Wales or those not exposed) did not affect the results. For the single landfill site analysis, the postcode method produced a significantly increased risk whereas the centroid method did not find a significant increased risk. Hence, an accurate population at risk in the exposed area is required to determine if a significant increased risk exists. The postcode method produced a non-significant decreased risk for female childhood cancer compared with a significant decreased risk for female childhood cancer using the centroid method. Even though it cannot be stated that the postcode method is better than the other two methods, since the true exposed region is not known, the population estimate at risk should provide more accurate results compared with the other two methods.

As with any analysis, there are various factors in this research that could have affected the results. Age-sex-deprivation standardisation was carried out but there could have

been other confounding factors which were not taken into account. The coordinates of the landfill sites were of the central position of each landfill site (or as close as possible to the centre). Differing results could have been obtained if the coordinates of the landfill gate were used. Extrapolation of the 1991 Census population figures at ward and ED level for the years 1982-2001 were calculated based on their respective local health board level age and sex distribution for their respective years. However, comparing these with using the 1991 Census figures only for the twenty-year period did not affect the results. Migrational changes may have occurred in specific wards or EDs of a local health board that would not have been taken into account, even after extrapolation. However, through the analysis shown here, very little effect on the SIRs was seen when exploring various options. It should be reiterated that this study was not an epidemiological one but one to compare the impact on results of differing methodologies.

Brain Cancer

The postcode method produced a SIR of 0.99 using wards and EDs as the geographical unit within 2km of landfill sites in Wales. Both results were non-significant for the initial analysis. Similar SIRs were obtained using the intersection method and centroids method but the number of expected cases in the analysis ranged from 1436.83 using the centroids method to 3764.90 using the intersection method (postcode method = 1544.44 expected cases) for ward analysis. This gives an idea of the varying population at risk used in each of the methods. The results quoted above used all Wales age-sex-deprivation specific rates to calculate the expected numbers of cases in the exposed population. However, when only those living greater than 2km from landfill sites were used in the analysis to calculate expected figures, this figure only increased slightly resulting in very similar SIR as those quoted. The "unexposed" population at risk was defined further as those living greater than 2km from landfill sites and additionally, those living within 2km of landfill sites whose diagnosis date was not within the operation dates of the landfill sites. This cut the number of observed cases by approximately a half. Since this is an ecological study, this result is exploratory rather than confirmatory as other factors could have caused this result. For electric power line analysis, 105 cases were observed within 100

metres of an electric power line. No significant results were found. Analysis was extended to within 1000 metres of power lines to enable a larger population at risk; however, no significant results were found although the SIR gradually approached unity to within 1000 metres of power lines in Wales.

The focused space time scan statistic produced a most likely cluster of radius 4.8km for the three year period 1992-1994. When this cluster was analysed using the postcode method, a similar SIR was obtained (compared with the centroids method that the spatial scan statistic uses) of 1.65. Taking the latency period into account increased the SIR to 1.67 using the unexposed population as all Wales and increased further to 1.72 when the unexposed population was those living greater than 2km from the landfill site. These results were significant.

Leukaemia

The postcode method produced SIRs of 0.99 and 1.00 using wards and EDs respectively for the initial analysis. The other methods produced similar SIRs, but as before, the observed and expected number of cases in the analysis varied greatly. Calculating expected figures based on those not exposed to the population produced a slightly lower SIR of 0.99. Taking into account the operational time of the landfill sites and only those diagnosed during operation gave the same SIR of 1.01 (non-significant) and halved the number of observed and expected number of cases in the analysis. Electric power line analysis produced significant increased risks of leukaemia for those living within 400 metres up to within 1000 metres of power lines in Wales. This result cannot confirm the link between power lines and leukaemia since other factors may have affected the result.

A most likely cluster of radius 42.4km was found in West Wales for the period 1996-1998 when using the space time scan statistic. The postcode method produced a slightly lower SIR but was still significantly increased. However, when taking into account the latency period, due to the landfill site being in operation years before the cluster period, only those cases greater than or equal to 10 years of age and diagnosed in 1996 were

included in this analysis. This resulted in a significant result for males and produced a non significant result for females. The latency period accounted for a decrease of 69% in the number of observed cases (from 90 to 28) and hence slightly wider confidence intervals.

Childhood Cancer

Similar SIRs were obtained for childhood cancer (as with the previous cancers) at 1.02 for wards and EDs using the postcode method. The numbers of expected cases in the analysis were more than double when using the intersection method compared with the postcode method. This figure increased by another 3% to 1.05 when taking into account the opening times of the landfill sites and the number of observed and expected cases. However, all results were non-significant. Significant decreased risks were found for females living within distances of 100 metres to within 400 metres of electric power lines. This significant result was not apparent for males and females combined.

The spatial scan statistics can be used as four separate methods (local spatial, focused spatial, local space-time, focused space-time) and can aid the user in identifying specific areas of interest that each of the methods have in common in the identification of the various most likely clusters. Only the focused space-time scan was examined in this area of research. These areas can be investigated further to determine possible reasons as to why some methods gave very similar clusters.

Overall summary

Using the centroid method, the exposed population at risk contains some persons that are not at risk and the unexposed population contains some at risk using the centroid method. This measure tends to bias the estimate of risk towards one. However, if there is no effect, as seems to be the case here, it is irrelevant. In general, the postcode method should be used over the intersection method and centroids method since, although some SIRs were very similar for the other two methods, the numbers of observed and expected

cases varied widely between each of the methods due to the wide variation in the population at risk within the exposed region. Also, the intersection method and centroid method are more likely to give a poor estimate of the population at risk due to the aggregation method used to calculate the population at risk.

Other factors apart from environmental exposure not taken into account may have resulted in the figures in the analysis. Such factors not accounted for include quality of diagnosis of disease, classification of disease, accuracy of population estimates, variations due to chance, residential location only taken into account, migration of population and day to day movement such as place of work.

Multiple testing is also an important issue and should be taken into account. i.e. consider a test that has been conducted to examine whether an increased risk exists around a particular hazardous source. Suppose no increased risk was found. The analyst may decide to examine the same exposed region but use a different time period or different cancer to find a significant result. On average, 1 in 20 tests will be significant at the 5% level of significance. Various methods have been analysed, various cancer sites, various distances (for power lines). This is deemed multiple testing. To allow for this, the analyst may want to adjust the p-value obtained by using adjustments described in theme one. Alternatively, if calculating confidence intervals, 99% confidence intervals should be calculated as opposed to 95% confidence intervals.

Address history is another important factor. How long had the person lived in their address at diagnosis? Had they previously lived near another landfill site or in an unexposed population at risk? The prevailing wind direction was not taken into account which may have caused a non-circular area of risk around landfill sites, depending on the direction of pollution.

To summarise, it has been shown that results can vary depending on the spatial units under consideration or reference to calculate expected figures. However, the postcode method results showed general comparability when using different geographical units.

Initial investigation showed no evidence of increased risks within 2km of a landfill site in Wales. A slightly significant decreased risk was found for female childhood cancer for those living within 100m of electric power lines in Wales. When using the space-time scan method, areas of significant increased risks were located and these could be investigated further. It proved a useful tool in investigating areas of increased risks. Latency periods should be investigated to determine whether a case could have actually been exposed to a cancer causing agent near to where they reside. The effects of this can vary depending on the age distribution of the disease to be investigated and the years of operation of the landfill sites as was observed when examining the results of the focused space-time scan statistic, especially for leukaemia.

The postcode method gives a more accurate population at risk in the area to be analysed compared to the other methods studied here. Thus, a recommendation is that this method be used at cancer registries over other methods currently used and the expected numbers of cases in the exposed region should be calculated using only those "not exposed".

4. THEME THREE

Spatial Variations of Relative Survival in Wales

4.1. Aims and objectives

Cancer is an unavoidable part of life for many people. Approximately one in three men and one in four women in Wales will be diagnosed with cancer before their 75th birthday (WCISU, 2002). Better treatment and outcome in recent years has led to an improvement in survival from the majority of cancers. Although survival has improved, there may be important spatial variations of survival.

This area of research examines cancer survival for female breast cancer and colorectal cancer in Wales. These two cancer sites were chosen due to the large number of diagnoses and large numbers of deaths that are required to obtain reliable survival calculations at small levels of geography. In the past, there has not been a suitable geographical unit to calculate reliable relative survival rates at small area level. However, the ONS have recently defined super output areas from the 2001 Census: lower super output areas and middle super output areas (MSOA). Higher super output areas are still to be defined as of February 2008. Relative survival rates are explored in Wales, Local Health Boards (LHB) in Wales (22 in total) and Middle Super Output Areas (MSOA) in Wales (413 in total) for both cancers to determine whether survival from cancer shows any spatial patterns in Wales or if there are areas with significantly increased or decreased survival rates at small area level. This area of research aims to apply smoothing techniques to cancer survival rather than cancer incidence. There are very few, if any, past studies regarding smoothing survival rates at small area level. The reason for this is that in the past a 'suitable' geographical unit has not been available to analyse survival data. Smoothing survival rates "borrows strength" from neighbouring areas to determine whether any survival patterns exist.

Based on the initial research theme, Moran's statistic is used to investigate the autocorrelation of neighbouring survival rates (Oden's method is not used since the population has already been taken into account when calculating survival rates).

Unusually low or high cancer survival rates could be due to various factors such as prognosis, surgeon's expertise, effectiveness of screening programmes, stage of disease or distance to hospital, to name but a few.

To summarise, the following aims and objectives were set for this theme:

- To determine whether relative survival rates of female breast cancer and colorectal cancer differ by LHB in Wales.
- To examine relative survival rates of female breast cancer and colorectal cancer at a small geographical unit (MSOA) to determine any spatial patterns that may exist.
- To investigate areas of high and low relative survival rates via various smoothing models.
- To compare the smoothed models of relative survival with clusters located using the spatial scan statistic.
- To examine breast screening data at LHB level in Wales and to apply this to the female breast cancer smoothing model.

4.2. Background

4.2.1. Observed and Relative Survival

Survival analysis is concerned with the analysis of times to the occurrence of an "event". In cancer studies this is known as the time period between diagnosis and death for each patient. Cancer registries in the UK hold population based databases and follow up the patients held on this from diagnosis until death. Therefore, observational studies can provide the actual survival rates being achieved in the entire population and are a very important public health tool.

There are several approaches to estimating cancer survival in population studies. Considered here are observed (crude) survival and relative survival.

Observed survival is the probability of surviving all causes of death for a specified time interval. It is usually expressed as the percentage alive at a given time point (e.g. 1 year, 3 years or 5 years) since diagnosis. Problems with this method arise if comparisons are to be made between populations with different age distributions. Observed survival is likely to be lower in an older population since they are more likely to die not of the cancer, but from other causes.

Relative survival is the most widely used method in population studies. It is the ratio of the survival observed in the group of cancer patients to the survival that would be expected if they were subject to the same overall mortality rates by age, sex and calendar period as the general population. The expected probabilities are obtained from life tables for Wales that provide the life expectancy of persons for a given year by age and sex. The problems arising with crude survival are therefore overcome. It enables one to measure variations in cancer survival (or mortality) independently of variations in expected (background) mortality associated with age, geographic region, deprivation and calendar time.

Cancer survival is dependent on age at diagnosis and is in general likely to be lower in older patients. Therefore, if the age distribution of the general population at risk and cancer patients varies between different populations, comparing relative survival across these populations can be misleading. Age standardised relative survival overcomes this. Age and sex-specific relative survival rates are multiplied by the corresponding sex and group weight for a standard, reference population. These are summed to obtain a standardised rate. For Wales, the World Standard Cancer Patient Population (the proportion of cases in a particular age group and sex for a particular cancer) is generally used (Black et al., 1998). To obtain age-standardised relative survival rates at a lower level of geography in Wales (e.g. local health boards), it is more useful to use the proportion of cases in each age group by sex in Wales rather than use the World Standard Cancer Patient Population. i.e. a Wales Standard Cancer Population, due to the differing age distribution at diagnosis between other European countries.

4.2.2. Life Tables

A life table is a summary of mortality, survivorship and life expectancy for a specified population. In demography a complete life table is a mortality schedule showing detail for each single year of age and continuing until the last member of the cohort dies. Life tables were obtained from the Government Actuary Department (GAD)[6]. These tables are produced by sex and single year of age up to 100 and based on 3 years of data, e.g. 1980-1982, 1981-1983 up to 2003-2005. Each of the tables is based on the latest revised mid year population estimates and deaths data for a three year period.

Relative survival was computed using a STATA algorithm[7] based on the maximum likelihood method of Esteve et al (1990). As zero survival times are not accepted by STATA, a follow up duration of 1 day is imputed where the date of diagnosis is registered as the date of death on the WCISU database. This is a general rule that all cancer registries in the UK apply. The following time intervals (or break options) used by the London School of Hygiene and Tropical Medicine (LSHTM) are generally used to calculate relative survival rates in STATA: 1 month for the first 6 months; 3 months for the remainder of the first year; 6 months for the second year and yearly from 3 years until 5 years. For rare cancers and lower geography levels, fewer time intervals are used depending on the number of deaths in each interval and the 'event' that is being analysed. The usual set used by the LSHTM for rare diseases are 6 months for the first 3 years, then an interval of 2 years up to 5 years. In general, the time intervals are user defined but it is advised that there should be at least ten deaths per cell to enable a reliable survival calculation.

4.3. Literature Review

A major aim of cancer research is to improve the survival of cancer patients. It is thought that factors such as age at diagnosis, distance between home of the patient and treatment centre, diet, diagnostic factors and treatment given, socioeconomic factors (patients living in affluent areas tend to have better survival rates than those patients living in deprived

[6] Government Actuary's Department, London. http://www.gad.gov.uk/Life_Tables/Interim_life_tables.htm
[7] 'strel' command for estimation of relative survival written by Slogett A, Hills M, de Stavola B, Mander A. (1999).

areas) and place of residence affects cancer survival as shown in past studies. Few studies have investigated cancer survival at small area level due to the unstable estimates that can arise due to small numbers of cases in each geographical unit.

The following review is split into two sections: studies regarding cancer survival in general and studies that examine the spatial distribution (incidence, mortality and survival) of cancer. It is planned that both of these areas will be linked to investigate cancer survival on a spatial theme.

Cancer survival studies

Since there have been many published studies regarding cancer survival in general, the following studies have been selected for inclusion here based on the relevance to survival in the United Kingdom and various factors that could influence survival estimates.

It has been well documented that socio-economic deprivation is associated with decreased survival in patients with cancer (Lipworth et al., 1970). Mullee et al (2004) analysed 93687 breast cancer patients in England for the period 1992-1994, followed up to 31st December 1999 to enable the calculation of five-year relative survival. Analysis was conducted at health authority level (99 in total) and adjusted by various factors such as socio-economic deprivation (using the Carstairs index), mean age and race. Analysis showed that socio-economic and geographical indicators were the strongest predictors of the five-year relative survival rates. The mean five-year survival rate was 75%, ranging from 66% to 85% between health authorities. It was concluded that the significant variation in survival rates between health authorities could be partly explained by socio-economic status. Another possible explanation was the difference in health care between health authorities in relation to treatment guidelines and extent of initial investigations.

Coleman et al (2001) reported that deprivation was a major factor in the survival of cancer in England and Wales. 58 types of cancer were investigated for the diagnosis period 1971-1990 (followed up to December 31st 1995). Patients were allocated a

deprivation category (using the Carstairs index) via the construction of age-sex-deprivation life tables. Results show that for all cancers combined in the period 1986-1990, 12745 excess deaths from 492902 deaths would have been avoided if the deprivation categories other than affluent had experienced the same survival rates as the affluent group. Another study also concluded that deprivation was a major factor regarding survival (Wrigley et al., 2003). Gender was also found to be associated with all cause (all causes of death, as opposed to cancer cause specific deaths) survival. Hazard ratios ranged from a significant 18% increased risk for the deprived tertile of deprivation for univariate analysis to a significant increase of 15% for the deprived tertile of deprivation after adjustment of prognostic factors. Various studies in the USA such as those by Singh et al (2004) and O'Malley et al (2003) also concluded that socioeconomic status was an important factor in cancer survival and an important factor when monitoring trends in cancer survival.

Yu et al (2004) analysed survival measures for 25 major cancer sites in 17 health services in New South Wales, Australia for the diagnosis period 1991-1998. Region-specific risks of excess deaths due to cancer were estimated adjusting for age, sex and spread of disease at diagnosis. Empirical Bayes methods were used to shrink estimates of the region-specific risks of excess death due to cancer and found that 6.4% (2903 of 45047 deaths) of the deaths within 5 years could have been avoided if there was no regional variation in cancer survival by shifting the State average risk to the 20th centile of the distribution of region-specific risks of excess deaths. Other studies by Dickman et al (1997), Farrow et al (1996) and Twelves et al (2001) stress that place of residence is an important determinant of survival from cancer and is due to factors such as access to primary health care, diagnostic and treatment facilities. Dickman et al. (1997) stated that around 2.5% of all cancer deaths could be prevented by eliminating regional variation in cancer survival if everyone received the same level of care.

Another study by Yu et al in 2005 examined the impact of area of residence on colon and rectal cancer survival rates in New South Wales, Australia for the diagnosis period 1992-2000 followed up to 31st December 2001. Period analysis was undertaken. Period

survival focuses on a recent time interval e.g. 1996-2000 in which the patient's survival is followed up and excludes short-term survival of patients diagnosed before the start of the interval. i.e. 1992-1995. It calculates survival based on recent follow up information as opposed to the calendar year of diagnosis. The geography level used was Area Health Service (17 in New South Wales). Numbers of cases varied greatly between Area Health Service (e.g. from 80 rectal cases in the least populated area to 1799 rectal cases in the most populated area). Surgical experience was highlighted as a possible difference in survival rates between Area Health Services. i.e. patients have a better prognosis when treated by surgeons with a higher caseload and specialist expertise.

Treatment factors are also an important aspect to cancer survival. A study by Allgood et al (2006) showed that a surgeon's specialisation in management of screen-detected breast cancers was associated with longer survival. Welsh breast cancer data were used for this study for the period 1989-1997 with patients being followed up to 1999. Another study by Hebert-Croteau et al (2005) concluded that larger hospitals with increased patient volumes were associated with improved survival using data from 5 regions in Quebec, Canada for the period 1988-1994. The time taken to see a specialist was not taken into account and may have influenced the survival statistics.

Mammography screening tends to diagnose breast cancer tumours earlier than would have been expected. Thus, the tumour is less advanced than would have been without screening and survival improves by at least the time until it would have been diagnosed. More than 20 countries have introduced mammography screening programmes (Shapiro et al., 1998) and have shown benefits to mammography screening. A significant mortality reduction in breast cancer patients was found in a study by Tabar et al (2003). Antinnen et al (2006) examined the effect of a population-based screening programme by comparing tumours diagnosed during the pre-screening period (1977-1986) to those of the screening period (1987-1997). Survival of breast cancer was 7% higher in the screened group at 73% compared with the pre-screened group. The study by Anttinen showed a significant change to a more favourable stage distribution during the screening period compared with the pre-screening period. This would have been expected since

breast cancers are being diagnosed earlier than they would have been if the cancer had not been diagnosed via screening. However, there was no evidence to suggest that detection by screening was an independent prognostic factor with a hazard ratio of 0.75 and a non-significant p-value of 0.17. This is further explored in this area of research using screening figures obtained from Breast Test Wales from 1989 to 2000 by LHB in Wales.

There are also thought to be rural factors regarding survival from cancer. Campbell et al (1999) analysed 63796 patients who were diagnosed with various cancers. The study concluded that there was strong evidence to suggest that increasing distance from a cancer centre was associated with poorer survival reflected by more advanced stage of disease and less adjuvant therapy. It was thought that patients living far away were less likely to be diagnosed before they died or diagnosed at a later stage, especially for cancers such as stomach, breast and colorectal cancer, and therefore having lower survival rates (Launoy et al., 1992).

Spatial studies in cancer incidence, mortality and survival

The previous studies examine particular factors in relation to survival of cancer. The following studies examine the spatial distribution of cancer, in particular those that explore incidence, mortality and survival patterns.

Osnes et al (1999) proposed a method using a fully hierarchical Bayesian approach that incorporated spatial autocorrelation of hazard ratios using breast cancer and malignant melanoma in Norway. Municipalities (439 in Norway) were used as the geographical unit but some regions were scarcely populated and hence no cases were observed. It was found that there were areas in Norway of increased cancer survival for both cancer sites. Only clinical stage I breast cancer was analysed with Osnes; this was due to the greatest treatment gain for such patients and having the greatest geographical variation. However a misclassification of clinical stage between regions could have explained the regional differences due to the quality of radiological investigations. Another Norwegian study by

Kravdal (1998) suggested that the excess in mortality from malignant melanoma and breast cancer was higher in lower socioeconomic groups than in higher socioeconomic groups, thus the need to adjust for this factor.

Pascutto et al (2000) analysed incidence of laryngeal cancer in the Thames region of the UK for the period 1985-1993 to discuss the statistical issues involved with small area mapping. Inadequacies can arise due to errors in the numerator (under-registration) and denominator (under-enumeration at Census). Areas with small populations tend to have high sampling variability. Confounding should be taken into account to adjust rates and risks accordingly. Pascutto concluded that care should be taken when assigning prior distributions for hierarchical models since results can vary greatly.

Johnson (2004) used hierarchical Bayes spatial modelling techniques to produce maps of smoothed standardised incidence ratios (SIR) for incidence of prostate cancer for the diagnosis period 1994-1998 in New York State, USA. Johnson concluded that differences across the state may be attributed to socio-demographics and other risk factors. Johnson calculated SIRs by age and race. It was noted that for specific ZIP codes on the border of New York State, the less populated areas tended to have greater uncertainty due to there being fewer neighbours to "borrow" strength in calculations of smoothed SIRs. Difference in screening and a seasonal residence effect were also noted for the differences in SIRs between ZIP codes in New York State.

Very few studies have examined clustering techniques in relation to survival of cancer due to the very small numbers of deaths when looking at small area level survival analysis. A spatial scan statistic was proposed by Huang et al (2007a) based on an exponential model that could incorporate survival data. Survival of prostate cancer was analysed for the diagnosis period 1984-1995 in Connecticut, USA. The model was adjusted for potential confounding factors such as age, race/ethnicity and disease stage to locate areas of high or low survival. Randomly generated data were used to assess the power of the scan statistic. Comparing the results in this study with a previous study by Gregorio et al. (2004) regarding incidence, areas that showed greater than expected

incidence appeared to have better than expected survival. Conversely, areas that showed less than expected incidence appeared to have poorer survival. It was thought that this could be due to inadequate detection and/or treatment of cases.

A further study (Huang et al., 2007b) described the use of the cluster detection method using a spatial scan statistic based on an exponential survival model for colorectal cancer (stage III and stage IV disease) and lung cancer (stage I/II, III or IV) in the State of California and County of Los Angeles for the diagnosis period 1988 to 2002. Results showed potential for the clustering techniques by Huang et al and consistency between the cluster results and survival curves calculated by Kaplan-Meier methods.

4.4. Methods

All cancer registrations for female breast cancer (ICD 9 codes 1740-1749, ICD 10 codes C500-C509) and colorectal cancer (ICD 9 codes 1530-1549, ICD 10 codes C180-C199, C20, C21) were extracted from the WCISU database for the diagnosis period 1981-2000. These cancers were used due to the large number of cases in the analysis to enable robust results. Each case's postcode was updated with its current postcode using ProAddress, an extension in ArcGIS V8.3. Subsequently, the updated postcodes were allocated an easting and northing correct to one metre using the Ordnance Survey product CodePoint^{TM}. Each case was allocated a MSOA as defined by the Office for National Statistics (ONS) using the respective shape file in ArcGIS V8.3. There are 413 MSOAs in Wales with a minimum population of 5024 and a mean population of 7050 in Wales. The use of MSOA was simply to obtain a relevant geography level that contained sufficient numbers of deaths in individual areas. MSOAs were defined in 2001, whereas the data used was for the period 1981-2005 (twenty year diagnosis period where patients were followed up to the end of 2005). Ward level data would have produced very few deaths per geographical unit to enable accurate survival calculations.

The Welsh Assembly Government published the Welsh Index of Multiple Deprivation (WIMD) 2005 in 2005[8]. This is a measure of multiple deprivation calculated at small area level in Wales. WIMD 2005 was produced at Lower Super Output Area (LSOA) level. There are 1896 LSOAs in Wales with a minimum population of 1005 and a mean population of 1530 in Wales. The seven domains that make up the WIMD 2005 are income, employment, health, education (and skills and training), geographical access to services, housing and physical environment. Cancer registries in the UK use the income domain as the source of deprivation when calculating survival, thus this method will be adopted for all analysis presented here. The postcode of diagnosis for each case was assigned its respective WIMD score and its respective quintile of deprivation from most deprived (5) to affluent (1) based on the 1896 LSOAs in Wales. Each case was then allocated its respective MSOA. Note that WIMD 2005 is used in this theme to determine deprivation as opposed to Townsend in the previous two themes due to the geography level analysed in this theme.

Relative survival rates are usually quoted for the age bands between 15 years and 99 years since it is thought that childhood cancer survival rates differ to adult cancer survival rates. For this analysis, only 2 cases were observed for female breast cancer and 2 cases for colorectal cancer aged less than 15 years and diagnosed between 1981 and 2000 and were removed from the analysis. 53 cases were observed for female breast cancer aged over 99 years and 33 cases were observed for colorectal cancer aged over 99 years of age. These cases were also removed since it can be difficult to trace these patients by the ONS as alive or dead. All cancer registries exclude these age ranges.

The following lists the inclusion criteria for the relative survival calculations:

- Only primary malignant female breast or primary malignant colorectal cancers were analysed.
- Only cases aged 15-99 years were included in the analysis.

[8] http://www.wales.gov.uk/keypubstatisticsforwales/wimd2005.htm

- Patients were followed-up until 31st December 2005, thus if a patient died after this date their status would be "alive" at 31st December 2005 for relative survival calculations.
- True zero survival times (patients diagnosed on the same day that they died) were imputed as 1 day since STATA does not accept zero survival times).

The following lists the exclusion criteria for the relative survival calculations:

- Patients that were only registered with a death certificate (the true survival time is not known as a patient's date of death would have been registered as the date of diagnosis since only the death date is known).
- Patients at the end of follow up aged over 99 years and who could not be traced by the ONS (since it is not known whether these patients are alive or dead).
- Cases that could not be assigned a LSOA (resulting in a MSOA, and therefore a quintile of deprivation) due to an unknown or incorrect postcode at diagnosis.

Analysis of relative survival in Wales and by LHB in Wales used two life tables to enable the calculation of expected survival based on the background mortality – 1985-1987 life table for cases diagnosed in the period 1981-1990 and 1995-1997 life table for cases diagnosed in the period 1991-2000 i.e. the mid-intervals (or as close as possible) of each of the time periods.

The age bands 15-49, 50-64 and 65-99 years were used to calculate age-specific survival rates for female breast cancer. These age bands were used since the screening age of breast cancer is between 50 and 64 years and incidence is generally low prior to this age. The age bands 15-64, 65-74 and 75-99 years were used to calculate age-specific survival rates for colorectal cancer by LHB in Wales for the period 1981-2000. The age bands used for colorectal cancer differ to female breast cancer due to the differing age distribution of colorectal cancer – it is generally diagnosed later in life. The age-specific survival rates were used to obtain an age-standardised rate by applying the age specific survival rates to the proportion of cases of female breast cancer or colorectal cancer in the specific age group. Relative survival figures were also calculated for the age band 15-99

(not age standardised) to determine whether there was a difference between the non-standardised and the standardised survival rates for each LHB in Wales.

4.5. Cancer datasets

Table 4.1 shows the distribution of cancer cases that were included in the relative survival calculations along with the proportion in each age category by LHB in Wales for female breast cancer and colorectal cancer for the period 1981-2000 respectively.

	Female Breast Cancer						Colorectal Cancer					
	Numbers of cases			Proportion of cases			Numbers of cases			Proportion of cases		
	15-49	50-64	65-99	15-49	50-64	65-99	15-64	65-74	75-99	15-64	65-74	75-99
Anglesey	151	282	391	18.3%	34.2%	47.5%	230	258	292	29.5%	33.1%	37.4%
Gwynedd	282	492	749	18.5%	32.3%	49.2%	416	457	595	28.3%	31.1%	40.5%
Conwy	243	509	883	14.9%	31.1%	54.0%	328	524	789	20.0%	31.9%	48.1%
Denbighshire	192	394	692	15.0%	30.8%	54.1%	297	396	573	23.5%	31.3%	45.3%
Flintshire	332	548	640	21.8%	36.1%	42.1%	425	464	481	31.0%	33.9%	35.1%
Wrexham	249	478	642	18.2%	34.9%	46.9%	378	387	491	30.1%	30.8%	39.1%
Powys	252	507	650	17.9%	36.0%	46.1%	384	436	536	28.3%	32.2%	39.5%
Ceredigion	149	290	410	17.6%	34.2%	48.3%	214	261	281	28.3%	34.5%	37.2%
Pembrokeshire	251	467	625	18.7%	34.8%	46.5%	355	407	440	29.5%	33.9%	36.6%
Carmarthenshire	399	806	917	18.8%	38.0%	43.2%	539	698	791	26.6%	34.4%	39.0%
Swansea	529	903	1182	20.2%	34.5%	45.2%	736	838	984	28.8%	32.8%	38.5%
Neath Port Talbot	326	555	713	20.5%	34.8%	44.7%	465	554	558	29.5%	35.1%	35.4%
Bridgend	327	460	596	23.6%	33.3%	43.1%	436	397	456	33.8%	30.8%	35.4%
Vale of Glamorgan	272	434	548	21.7%	34.6%	43.7%	350	357	419	31.1%	31.7%	37.2%
Rhondda Cynon Taff	483	903	1094	19.5%	36.4%	44.1%	643	758	800	29.2%	34.4%	36.3%
Merthyr Tydfil	105	220	259	18.0%	37.7%	44.3%	229	221	173	36.8%	35.5%	27.8%
Caerphilly	359	585	698	21.9%	35.6%	42.5%	484	512	522	31.9%	33.7%	34.4%
Blaenau Gwent	136	237	346	18.9%	33.0%	48.1%	240	275	283	30.1%	34.5%	35.5%
Torfaen	201	371	391	20.9%	38.5%	40.6%	289	309	324	31.3%	33.5%	35.1%
Monmouthshire	196	350	396	20.8%	37.2%	42.0%	243	285	324	28.5%	33.5%	38.0%
Newport	266	461	580	20.4%	35.3%	44.4%	355	434	428	29.2%	35.7%	35.2%
Cardiff	636	1061	1323	21.1%	35.1%	43.8%	826	892	1034	30.0%	32.4%	37.6%
WALES	**6336**	**11313**	**14725**	**19.6%**	**34.9%**	**45.5%**	**8862**	**10120**	**11574**	**29.0%**	**33.1%**	**37.9%**

Table 4.1: Numbers of cases and proportion of cases included in the analysis for female breast cancer and colorectal cancer, 1981-2000.

The majority of the LHBs in Wales follow a similar distribution of age-specific cases to Wales for female breast cancer as shown in table 4.1. Areas in North Wales such as Conwy and Denbighshire showed approximately 5% less cases in the age band 15-49 years for female breast cancer and nearly 10% more cases diagnosed in the oldest age band 65-99 years compared with Wales as a whole. For the years 1981-2000 Conwy and

Denbighshire had the lowest proportion of population aged 15-49 and the highest proportion of population aged 65-99 thus resulting in the figures quoted.

A similar distribution can be seen in North Wales for colorectal cancers (9% less cases in Conwy in the age band 15-64 year olds compared to the rest of Wales) due to the proportions of population in the youngest and oldest age categories. This has an effect on the relative survival figures since survival depends on the age at diagnosis – younger patients tend to have a better prognosis compared with the elderly.

Table 4.2 shows the incidence and mortality rates (incidence for those cases included in the analysis and mortality for those who were included in the analysis and had died before 31/12/2005) for female breast cancer and colorectal cancer by LHB in Wales for the period 1981-2000. Crude rates and Wales age standardised rates are quoted. The Wales age standardised rate takes into account the differing age structure seen in LHBs in Wales compared with Wales as a whole. It can be seen that even after age standardisation, rates in North Wales tend to be higher for both incidence and mortality and for both cancer sites examined compared to other areas of Wales. The highest rates were observed in Conwy and Denbighshire even after age standardisation. This allows for the fact that an older population live in these areas, the rates are still high. The lowest age standardised rates for female breast cancer incidence and mortality were found in Blaenau Gwent and Newport respectively. Newport and Vale of Glamorgan had the lowest age standardised rates for colorectal incidence and mortality.

	Female Breast Cancer				Colorectal Cancer			
	Incidence		Mortality		Incidence		Mortality	
	CR	WASR	CR	WASR	CR	WASR	CR	WASR
Anglesey	144.5	141.4	83.0	80.9	70.9	69.4	54.5	53.2
Gwynedd	153.8	146.3	89.8	83.9	78.3	73.1	62.2	57.4
Conwy	172.8	152.0	101.0	84.0	93.8	73.9	77.2	59.0
Denbighshire	164.1	149.8	103.9	91.6	87.0	73.9	72.0	60.1
Flintshire	130.2	138.5	75.5	82.1	60.6	67.1	48.4	54.2
Wrexham	132.3	136.0	80.8	83.7	63.3	65.6	51.6	53.6
Powys	143.1	137.1	80.2	76.2	70.1	65.9	54.8	51.1
Ceredigion	146.4	143.9	85.7	83.3	67.7	64.8	52.8	50.2
Pembrokeshire	145.0	142.6	84.3	83.2	67.5	67.2	54.0	53.8
Carmarthenshire	147.2	139.3	86.2	80.3	73.7	68.3	61.3	56.5
Swansea	135.2	134.8	78.2	77.8	69.1	68.7	55.5	55.1
Neath Port Talbot	135.6	132.5	79.5	77.2	70.4	68.8	57.3	55.9
Bridgend	128.5	130.8	74.2	76.3	62.4	64.7	48.8	50.8
Vale of Glamorgan	129.2	132.7	72.0	74.8	60.5	62.9	46.3	48.4
Rhondda Cynon Taff	127.2	129.5	76.0	77.9	58.8	61.6	48.9	51.5
Merthyr Tydfil	119.6	120.5	72.9	73.7	66.9	68.3	56.3	57.7
Caerphilly	118.6	125.6	69.9	75.5	56.8	62.9	45.9	51.4
Blaenau Gwent	118.8	117.8	72.7	71.9	68.3	68.0	56.8	56.5
Torfaen	128.4	132.1	74.8	78.2	63.9	68.3	51.9	56.0
Monmouthshire	139.1	139.1	72.5	73.3	64.8	65.8	50.9	51.7
Newport	117.1	121.0	68.5	71.8	57.1	60.5	46.7	49.8
Cardiff	121.2	129.5	70.3	75.7	57.9	61.6	47.1	50.4
WALES	134.7	134.7	78.7	78.7	66.4	66.4	53.6	53.6

Table 4.2: Crude rates (CR) per 100,000 population and Wales age standardised rates (WASR) per 100,000 population for incidence and mortality of female breast cancer and colorectal cancer by LHB in Wales 1981-2000, ages 15-99.

4.6. Relative survival for female breast cancer and colorectal cancer

4.6.1. Relative survival rates in Wales

36,871 female breast cancer cases were entered into the analysis of which 32,378 were eligible for relative survival calculations (4493 cases were excluded – the vast majority of the exclusions were due to the female breast cancer not being the primary cancer). For colorectal cancer, 6527 cases were excluded resulting in 30,556 cases being eligible for relative survival calculations.

Table 4.3 and table 4.4 show observed and relative survival rates for the period 1981-2000 for female breast cancer and colorectal cancer in Wales. The numbers of deaths for each calculation are also shown. For example, for breast cancer, the 3 year observed survival calculation was 70.25% for the period 1981-2000. This figure was based on 9498 deaths that occurred between 0 and 3 years from diagnosis. Observed and relative survival was steadily increasing throughout the twenty year period.

Even though there were a similar number of female breast cancer and colorectal cancer cases included in the analysis there were many more deaths from colorectal cancer than female breast cancer. For female breast cancer 5026 deaths (40%) from a total of 12,498 deaths occurred within one year of diagnosis. This figure rose to 60% (12,599 deaths) for colorectal cancer within one year.

It should be noted that relative survival rates were slightly higher than the observed survival rates. This was due to the expected survival of these patients having similar mortality rates as those of the general population. Just over 2 in 5 patients survived after 5 years from diagnosis of colorectal cancer for the period 1991-2000 whereas over 3 in 4 patients survived female breast cancer for the same period.

Observed	Female Breast Cancer			Colorectal Cancer		
Survival	1981-2000	1981-1990	1991-2000	1981-2000	1981-1990	1991-2000
1 year	84.04	79.85	87.18	55.75	50.55	60.14
95% CI	(83.63-84.44)	(79.16-80.51)	(86.68-87.66)	(55.18-56.32)	(49.70-51.39)	(59.37-60.90)
deaths (0-1yr)	5026	2713	2313	12599	6414	6185
3 year	70.25	64.55	74.54	37.54	33.06	41.36
95% CI	(69.75-70.75)	(63.74-65.34)	(73.91-75.17)	(37.00-38.08)	(32.29-33.83)	(40.61-42.11)
deaths (0-3yrs)	9498	4856	4642	18361	9014	9347
5 year	61.02	54.36	66.04	30.05	26.32	33.24
95% CI	(60.48-61.55)	(53.52-55.18)	(65.35-66.72)	(29.54-30.56)	(25.60-27.03)	(32.53-33.96)
deaths (0-5yrs)	12498	6287	6211	20774	10039	10735

Table 4.3: Observed survival in Wales 1981-2000.

Relative Survival	Female Breast Cancer			Colorectal Cancer		
	1981-2000	1981-1990	1991-2000	1981-2000	1981-1990	1991-2000
1 year	87.89	83.38	91.31	59.33	53.83	63.95
95% CI	(87.47-88.29)	(82.67-84.06)	(90.81-91.78)	(58.72-59.93)	(52.93-54.72)	(63.14-64.75)
3 year	77.48	70.96	82.42	43.82	38.87	47.99
95% CI	(76.94-78.00)	(70.09-71.80)	(81.75-83.06)	(43.19-44.44)	(37.97-39.77)	(47.12-48.84)
5 year	71.02	62.99	77.10	38.80	34.28	42.61
95% CI	(70.43-71.60)	(62.06-63.91)	(76.35-77.83)	(38.16-39.44)	(33.37-35.19)	(41.72-43.49)

Table 4.4: Relative Survival in Wales 1981-2000.

4.6.2. Relative survival rates by Local Health Board in Wales

Figure 4.1 and figure 4.2 show a comparison of five year relative survival rates by LHB in Wales for female breast cancer and colorectal cancer for the period 1981-2000 from lowest age standardised relative survival to highest age standardised relative survival (red for female breast cancer, brown for colorectal cancer). Unadjusted age standardised relative survival rates are also shown (pink for female breast cancer, orange for colorectal cancer). Age standardised relative survival rates use the age distribution of cases in Wales as a whole as opposed to the population distribution of Wales for incidence and mortality standardisation methods. It can be seen that the difference between relative survival figures for adjusted and unadjusted age standardisation was small for each LHB in Wales.

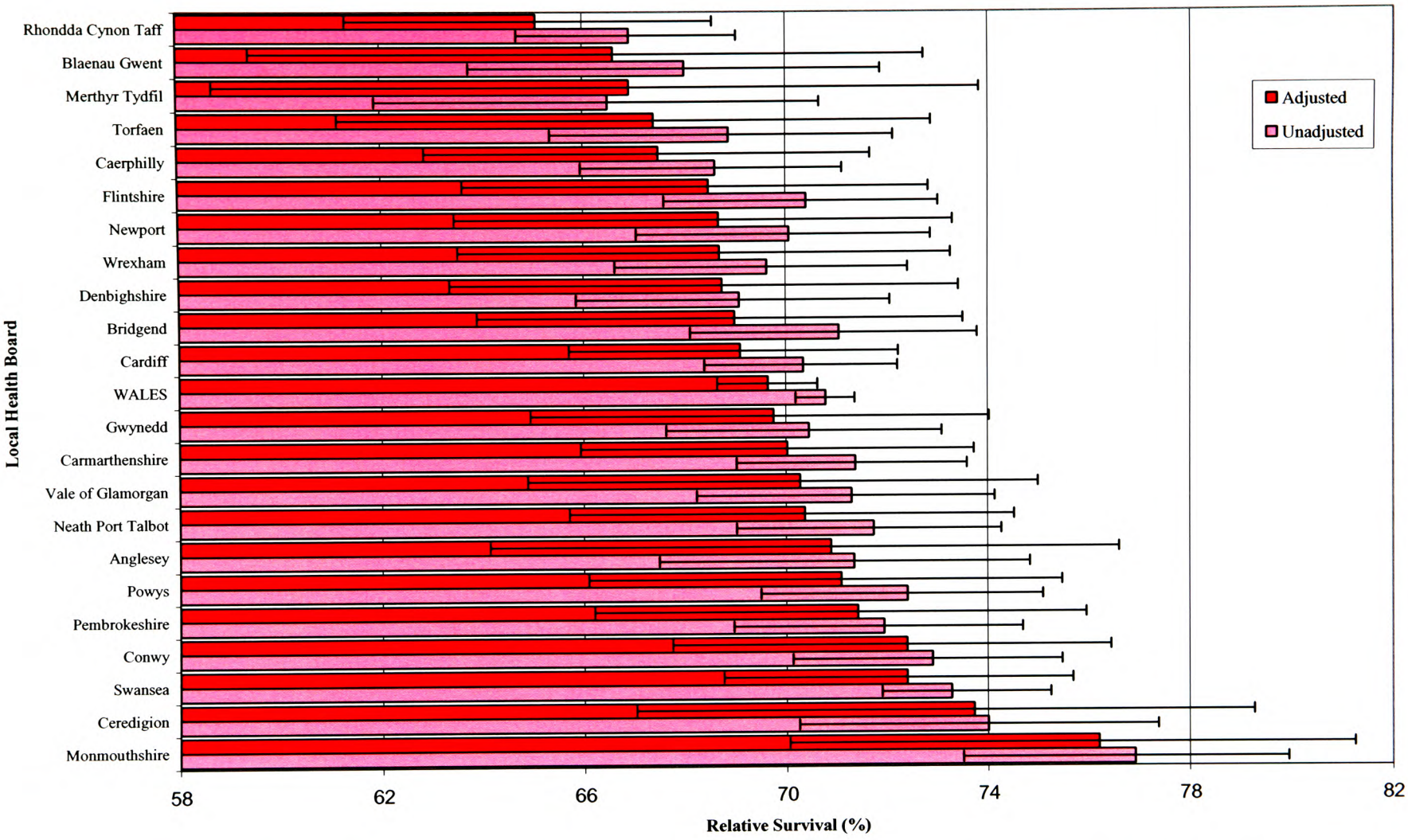


Figure 4.1: Five year relative survival for female breast cancer by local health board in Wales 1981-2000.

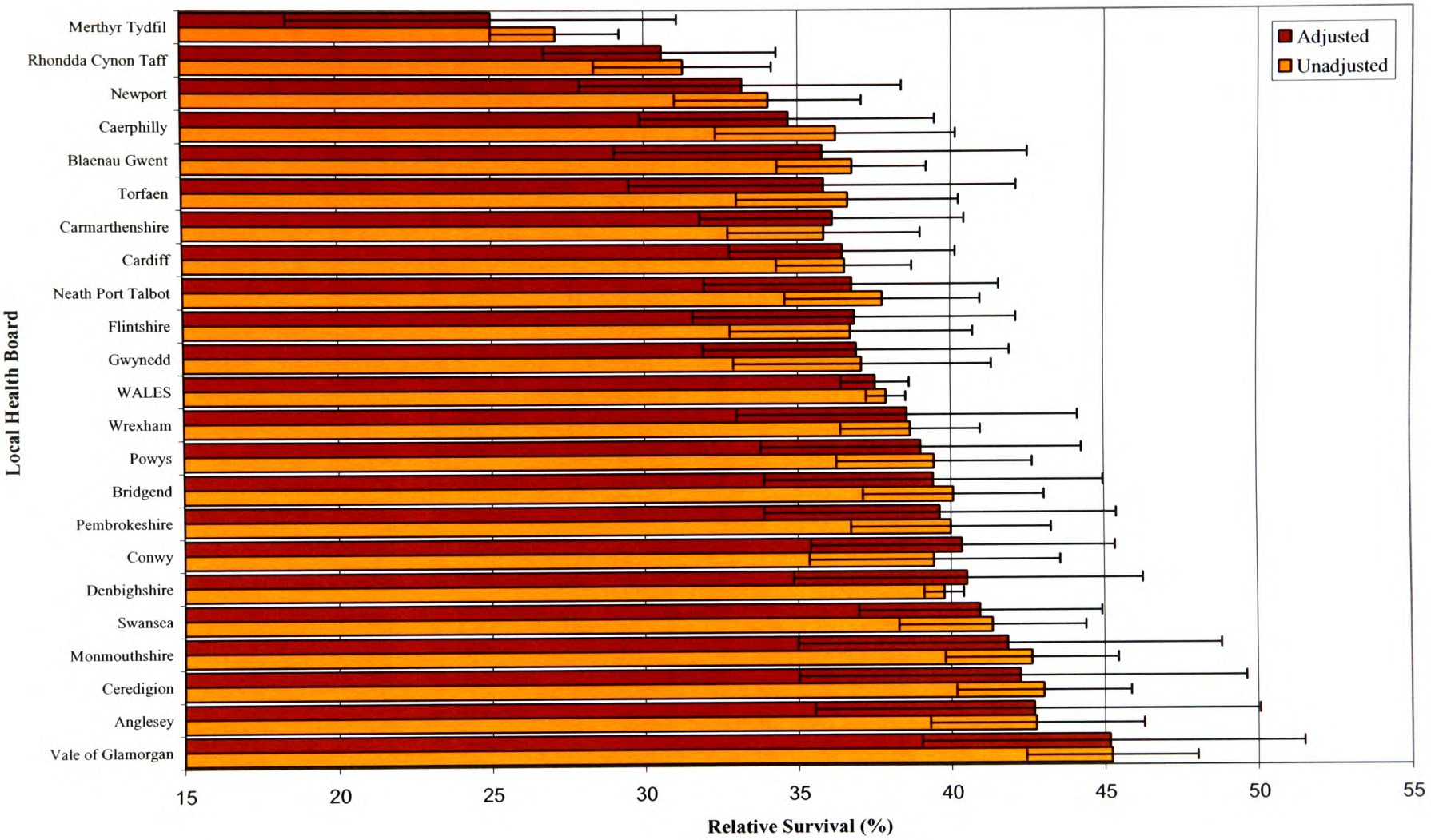


Figure 4.2: Five year relative survival for colorectal cancer by local health board in Wales 1981-2000.

LHBs in the South Wales Valleys such as Rhondda Cynon Taff and Merthyr Tydfil have the lowest relative survival for both cancers studied suggesting that there may be a socioeconomic factor in the relative survival of cancer. Affluent areas such as Vale of Glamorgan have the best relative survival for colorectal cancer. There is a wide variation in age standardised relative survival between LHBs in Wales. Five year age standardised relative survival rates varied between 67% (Rhondda Cynon Taff) and 77% (Monmouthshire) for female breast cancer and between 27% (Merthyr Tydfil) and 45% (Vale of Glamorgan) for colorectal cancer. When comparing between LHBs, there is a difference of over 10% for female breast cancer and around 18% for colorectal cancer survival with some LHBs displaying significantly different results to other LHBs. For example, Merthyr Tydfil (worst survival rates) had significantly lower five year relative survival rates compared with the Vale of Glamorgan (highest survival rates) for colorectal cancer.

Statistically significant differences in survival were found for colon and rectal cancer after adjusting for demographic factors by Yu et al., similar to results found above for colorectal cancer. Mullee et al found five year survival rates for female breast cancer in England varied widely between health authorities, again consistent with the results for female breast cancer in Wales. However, Mullee claims that this is due to differences in deprivation.

Figure 4.3 shows the relative survival rates by LHB for female breast cancer and colorectal cancer for the period 1981-2000 using a map of Wales to identify if neighbouring areas have similar survival rates.

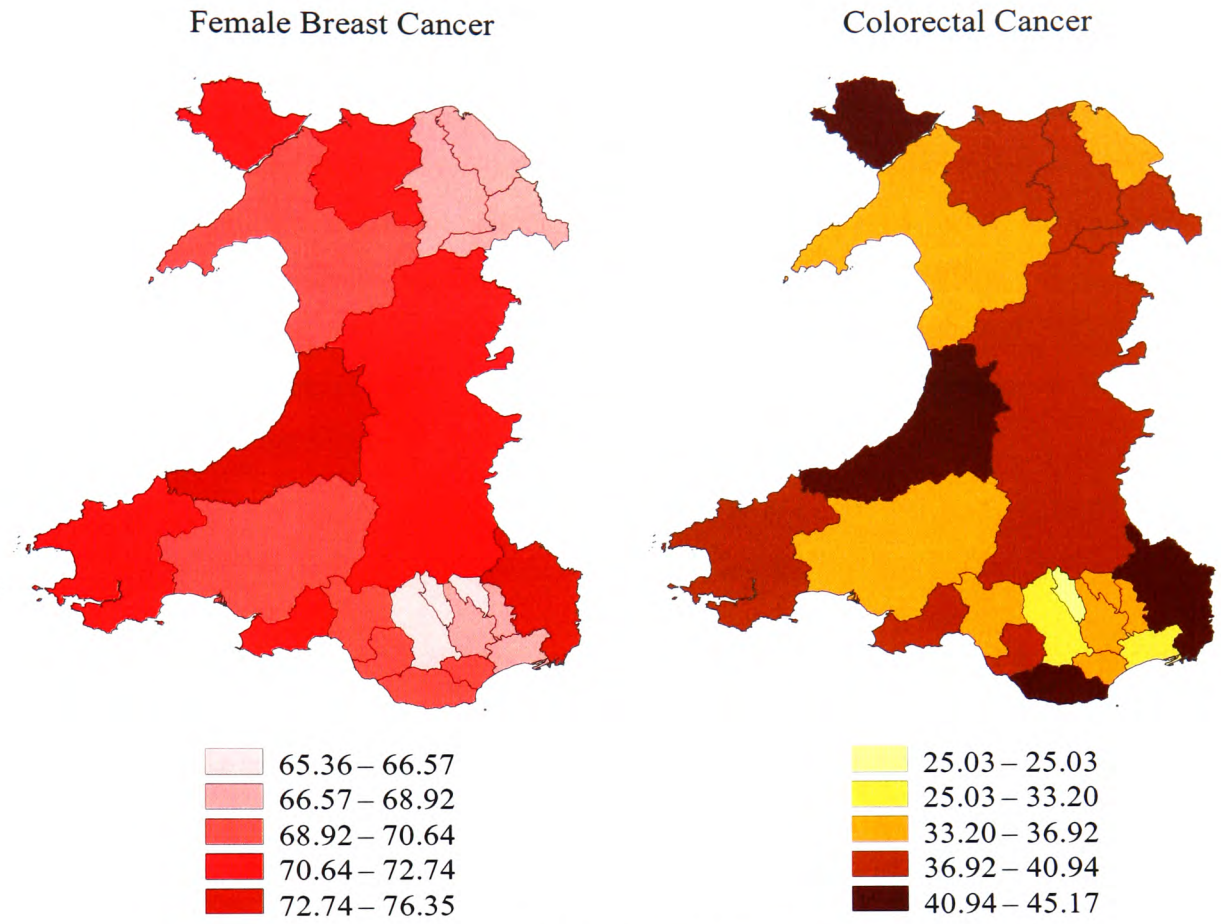


Figure 4.3: Relative survival rates by LHB in Wales, 1981-2000.

The break option used in this figure is called "natural breaks" and is the default classification in ArcGIS. Unless otherwise stated, all further breaks are based on this option. This method aims to minimise the sum of the variance within each of the classes. This method aims at identifying patterns, if any are present in the data.

The maps for female breast cancer and colorectal cancer for the period 1981-2000 show that the lowest survival rates for both cancers were located in South East Wales. Large differences can be seen for some neighbouring local health boards. This warrants a closer inspection of survival rates at a smaller geographical level.

4.6.3. Relative survival by MSOA in Wales

It was clear from the previous analysis that there were varying survival rates between neighbouring LHBs in Wales, approaching a 20% difference between Merthyr Tydfil and Vale of Glamorgan for colorectal cancer. In this section, relative survival is examined at a much lower level, MSOA to identify areas of high and low survival. Age, sex and deprivation were taken into account. To identify the areas of high and low survival, smoothing models were examined since there tends to be random variation at small area level due to small numbers.

Method

To enable reliable survival calculations, at least 10 deaths are required per geographical unit (Trent Cancer Registry, 2005), thus the MSOA level was used. A geographical unit smaller than this e.g. wards, would have too few deaths per area to calculate reliable survival statistics. The datasets for the survival calculations are for the periods 1981-2000, 1981-1990 and 1991-2000 with all patients followed up until 31st December 2005 to enable five year relative survival calculations for all patients.

Life tables are only obtainable from the Government Actuary Department (GAD) website by sex and single age. Due to the variation of relative survival by LHB in Wales, deprivation should be taken into account in the analysis of relative survival; hence the background mortality rates for Wales are required by single year of age, sex and deprivation. The GAD do not publish life tables by deprivation. Thus the life tables required for the analysis by MSOA were generated using data held at the WCISU.

Population figures for Wales were obtained for the periods 1981-1990 and 1991-2000 by single year of age obtained at the WCISU from the ONS from 0 to 84 years and over 85 years of age. All cause mortality figures were also obtained using data from the WCISU by sex, single year of age (from 0 to 100 years) and deprivation quintile. These figures were back calculated to obtain population figures for ages 85 to 100 by single year of age.

Secondly, population figures by LSOA were obtained by sex and five year age-band from East Midlands Public Health Observatory (EMPHO) for the Association of Public Health Observatories in collaboration with the UK Association of Cancer Registries (UKACR). These figures were aggregated to MSOA level and assigned their corresponding quintile of deprivation, from affluent (1) to most deprived (5) for the 413 MSOA with an equal number of MSOA in each quintile using the WIMD 2005. These figures were aggregated to obtain all Wales figures by quintile of deprivation and the proportion of the population was calculated in each quintile of deprivation for each five year age band.

These proportions by five-year age band were applied to the corresponding Wales population figures for 1981-1990 and 1991-2000 to obtain population figures by sex, single year of age and quintile of deprivation. Note that the same rate was applied to the five single year ages for the corresponding five year age band proportion since LSOA population figures were not available by single year of age.

Finally, the background mortality rate was calculated by sex, single year of age and quintile of deprivation by applying the calculated population figures to the corresponding background mortality figures from the ONS and the life tables generated for the periods 1981-1990 and 1991-2000 to obtain all cause mortality rates by sex, single year of age and deprivation.

Results

Table 4.5 shows the number of MSOAs where the number of deaths were less than 10. Relative survival calculations based on deaths less than 10 are considered to be unreliable. This rule can be overlooked when looking at spatial smoothing since analysis uses neighbouring rates to smooth the data. However as can be seen from table 4.5, there are very few MSOAs with deaths less than 10 for colorectal cancer. The numbers of MSOAs with deaths less than 10 for female breast cancer are higher if looking at the two ten year periods individually rather than the twenty year period 1981-2000. Thus, caution

is advised when interpreting results for female breast cancer for the ten year periods 1981-1990 and 1991-2000.

Deaths	Female Breast Cancer			Colorectal Cancer				
	1981-2000	1981-1990	1991-2000	1981-2000	1981-1990	1991-2000	Males	Females
0	0	1	0	0	0	0	0	0
1	0	1	0	0	1	0	0	1
2	0	1	1	0	3	0	0	0
3	1	1	2	0	0	1	1	2
4	0	5	1	1	2	1	1	1
5	0	5	10	0	0	0	2	2
6	0	8	8	0	1	1	1	1
7	0	13	9	0	4	2	0	2
8	3	23	20	1	3	1	2	3
9	1	21	14	1	6	5	2	4
>=10	408	334	348	410	393	402	404	397

Table 4.5: Number of MSOAs with less than 10 deaths per MSOA.

Figure 4.4 shows five year relative survival rates by MSOA in Wales for female breast cancer for the three periods 1981-2000, 1981-1990 and 1991-2000.

The three maps in figure 4.4 are presented with the same equal intervals. The total number of deaths within 5 years of diagnosis falls slightly in the latest ten year period (6205 deaths) compared with the period 1981-1990 (6285 deaths). However, there has been a clear increase in survival for the later period 1991-2000 compared with the earlier period. Since the number of deaths has remained relatively stable between the two periods, the increase in survival that is seen is due to more cases diagnosed in the later period compared with the earlier period (a 28% increase compared with the earlier period).

For the period 1981-2000, the majority of Wales had a five year relative survival rate between 62.94% and 80.53% (light brown). The average five year relative survival rates of all MSOAs for the three periods were 70.51% with range 49.00% to 89.75% for 1981-2000, 62.03% with range 10.18% to 94.98% for 1981-1990 and 76.92% with range 43.15% to 98.12% for 1991-2000. There was a very large range for the ten year period

1981-1990 at nearly 85%. There are localised areas of high survival (dark brown) in MSOAs throughout Wales. Examining the ten year periods, there has been a definite improvement in five year relative survival due to the dark colours seen in the period 1991-2000 compared with 1981-1990. The high relative survival rates in 1991-2000 are seen in areas of North Wales and localised areas in West and South Wales. However, as noted earlier for female breast cancer the ten year period 1981-1990 has 79 MSOAs with less than 10 deaths while the period 1991-2000 has 65 MSOAs with less than 10 deaths. Thus caution is advised in the interpretation of these results.

Relative survival for colorectal cancer were analysed by time period (1981-2000, 1981-1990 and 1991-2000) and by sex and are shown in figure 4.5 and figure 4.6.

The total number of deaths within 5 years of diagnosis increased in the latest ten year period (10,735 deaths) compared with the period 1981-1990 (10,039 deaths), a similar pattern was seen for female breast cancer. Again, incidence has increased in the later period resulting in higher survival rates compared with the earlier period.

The first noticeable difference between the female breast cancer maps and the colorectal cancer maps in figure 4.5 is the variation seen for neighbouring survival rates for colorectal cancer between the three periods, whereas female breast cancer appears to show similar neighbouring survival rates. The average five year relative survival rates of the MSOAs were 32.10% with range 6.77% to 64.67% for 1981-2000, 27.06% with range 0.02% to 71.16% for 1981-1990 and 36.76% with range 4.17% to 74.37% for 1991-2000. Like female breast cancer, there appears to have been an improvement in survival for the period 1991-2000 compared with the period 1981-1990 due to the darker colours. There were no apparent areas of high and low relative survival rates from the maps although there were a few MSOAs in mid-Wales that appeared to have low relative survival rates, perhaps an indication of poor access to treatment at hospitals. Figure 4.6 shows the corresponding colorectal maps for persons, males and females for the twenty year period 1981-2000.

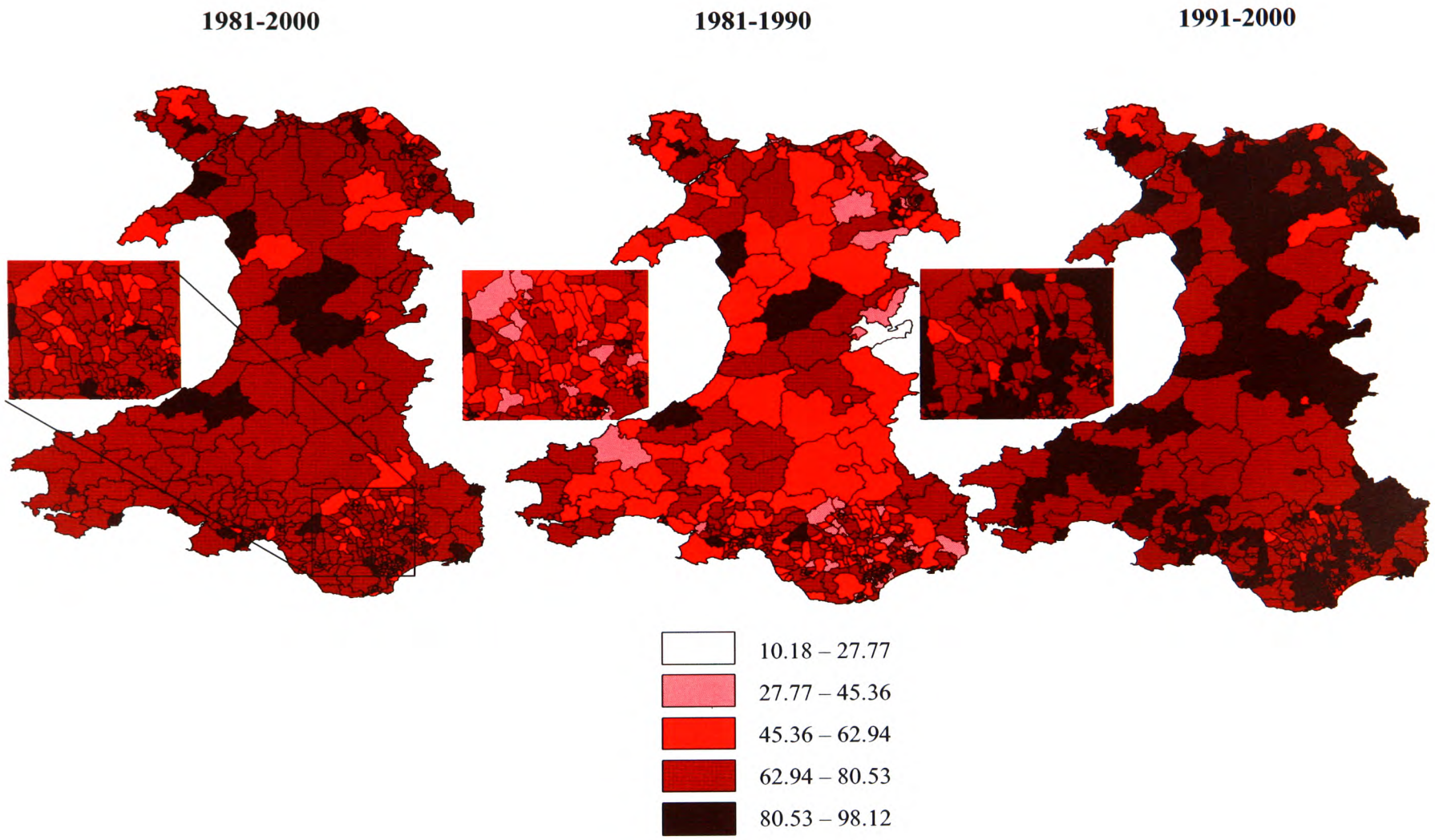


Figure 4.4: Five year relative survival for female breast cancer by MSOA in Wales.

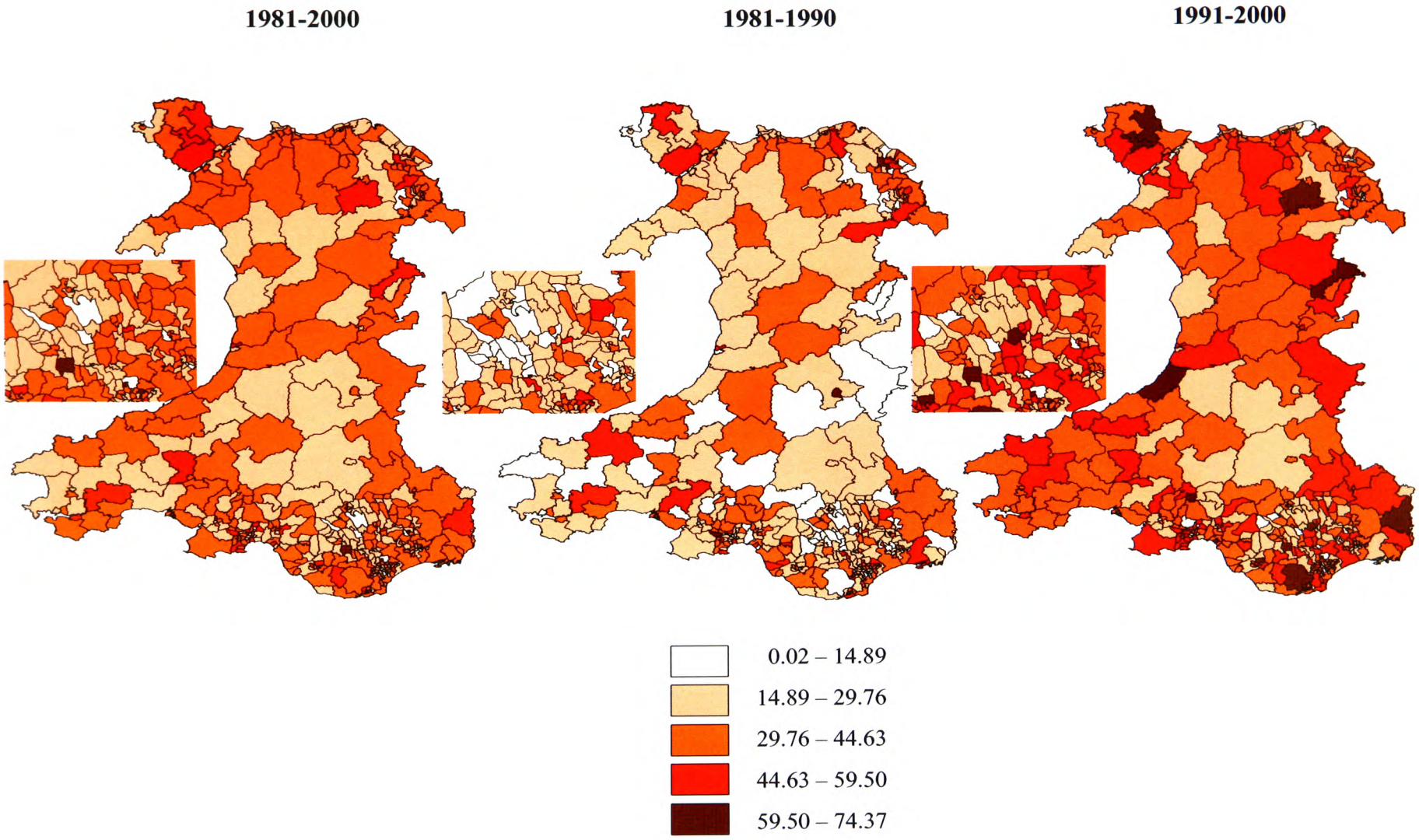


Figure 4.5: Five year relative survival for colorectal cancer by period in Wales 1981-2000.

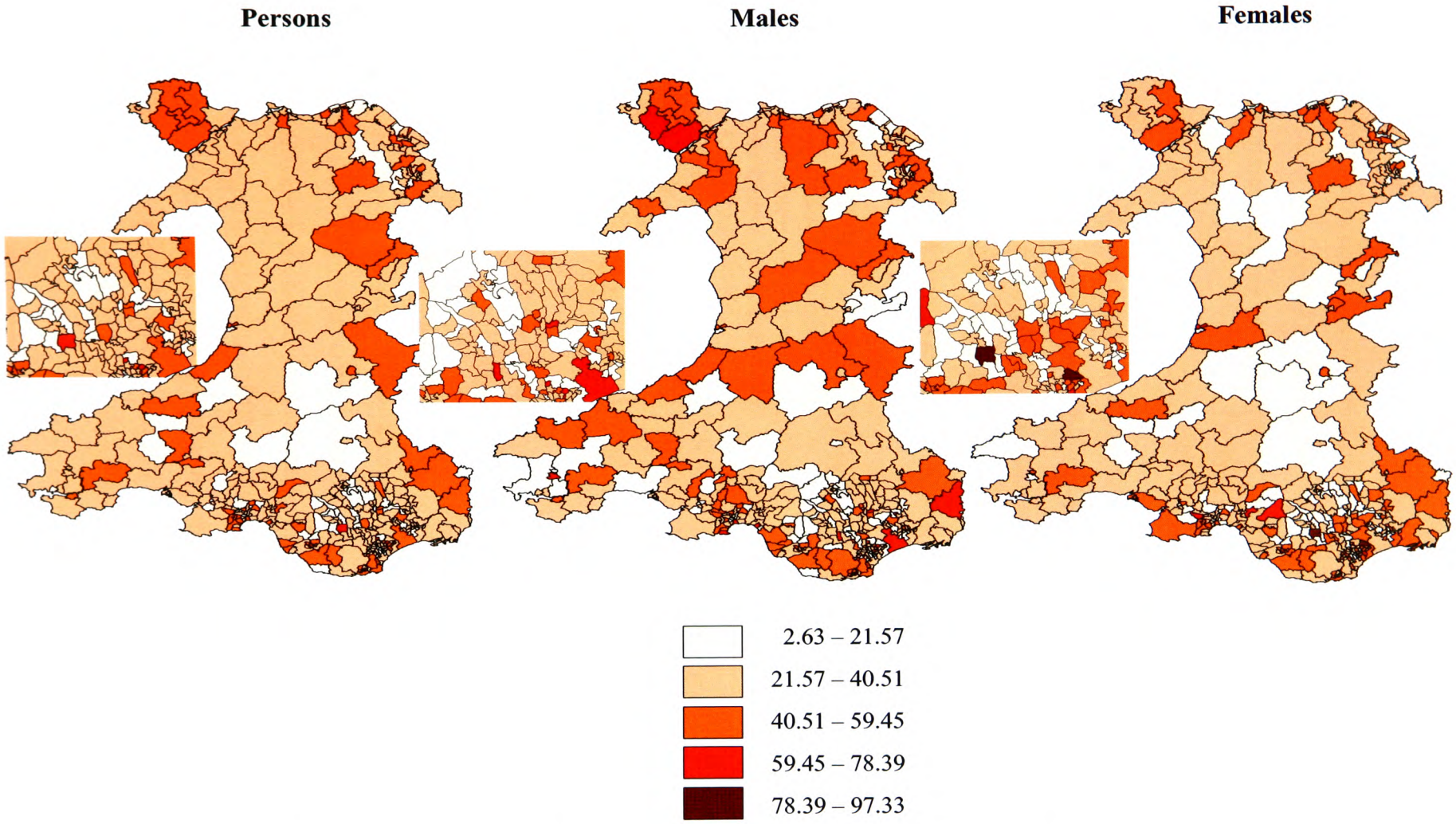


Figure 4.6: Five year relative survival for colorectal cancer by sex in Wales 1981-2000.

Examining all three maps in figure 4.6, there appears to be little difference in the five year relative survival rates between MSOAs by sex since the three maps appear very similar on first inspection. This is due to the average five year relative survival rates of all MSOAs being very similar at 32.10% for all persons (ranging from 6.77% to 64.67%), 33.08% for males (ranging from 2.63% to 74.72%) and 31.76% for females (ranging from 4.10% to 97.33%).

It was thought that for colorectal cancer, survival would be lower in rural areas compared with urban areas since patients would have to travel further to hospitals and thus cancers would be far more advanced at diagnosis. To test this, the urban/rural status was obtained from ONS[9] for all MSOA. Table 4.6 shows the five year relative survival rates and 95% confidence intervals for each of the cancer datasets for urban and rural areas.

Cancer	Urban	95% CI	Rural	95% CI
Female Breast Cancer, 1981-2000	70.52	(69.77, 71.25)	72.26	(71.27, 73.21)
Female Breast Cancer, 1981-1990	62.43	(61.26, 63.57)	64.15	(62.57, 65.68)
Female Breast Cancer, 1991-2000	76.82	(75.87, 77.74)	78.18	(76.94, 79.36)
Colorectal Cancer, 1981-2000	38.52	(37.73, 39.32)	40.10	(39.01, 41.18)
Colorectal Cancer, 1981-1990	34.48	(33.35, 35.61)	34.22	(32.69, 35.76)
Colorectal Cancer, 1991-2000	41.97	(40.86, 43.09)	45.01	(43.49, 46.51)
Male Colorectal Cancer, 1981-2000	38.57	(37.46, 39.69)	41.98	(40.44, 43.51)
Female Colorectal Cancer, 1981-2000	38.36	(37.22, 39.50)	38.04	(36.50, 39.57)

Table 4.6: Five year relative survival statistics for urban and rural areas in Wales.

There were nearly twice as many urban areas (269 MSOAs) as rural areas (144 MSOAs) in Wales. However, the rural MSOAs covered a much larger area in Wales than the urban MSOAs but the urban areas accounted for a larger population. In general, it appeared that there were higher survival rates in rural areas compared with urban areas. There was a marginally significant difference in the survival rates for female breast 1981-2000, colorectal cancer 1991-2000 and male colorectal cancer 1981-2000. It appeared that for colorectal cancer for the period 1991-2000, the significance was borderline and probably due to male colorectal cancer from the figures quoted in table 4.6.

[9] http://www.statistics.gov.uk/geography/nrudp.asp

Campbell et al found strong evidence that increasing distance from a cancer centre was associated with poorer survival, especially for breast and colorectal cancer. This does not seem to be the case regarding the results in table 4.6 since the higher survival rates appear to be in the rural areas.

4.6.4. Smoothing techniques for relative survival by MSOA in Wales

The numbers of deaths per MSOA were less than 10 for a small number of MSOAs. Survival estimates are unreliable when there are few deaths in a geographical unit. If this is the case, it is unclear whether there are any similar survival patterns in neighbouring areas. To overcome this, spatial smoothing is explored. This method borrows strength from neighbouring areas to examine survival patterns in the data.

Windows version of Bayesian inference Using Gibbs Sampler (WinBUGS)[10] is freely downloadable software that can be used for Bayesian analysis. A model is specified in the WinBUGS language along with any prior information to update the probability that a hypothesis may be true. Observed survival rates and expected survival rates for each of the MSOAs were entered into the Normal model (see Appendix D for the model used) along with a matrix of MSOA nearest neighbours for each MSOA in Wales for analysis. The survival rates used were those that were obtained when using the maximum likelihood method using the 'strel' command in STATA as previously described. The aim of using WinBUGS is that smoothed relative survival rates will be produced by MSOA that take into account each MSOA's neighbouring relative survival rates. It is hoped that patterns of high and low survival are observed that would not have been apparent pre-smoothing.

A burn in period of 10,000 iterations was used in WinBUGS before analysing the data (i.e. the initial number of iterations to discard). The analysis was based on a subsequent 20,000 iterations. Three periods were used in the analysis; those being 1981-2000, 1981-

[10] http://www.mrc-bsu.cam.ac.uk/bugs/

1990 and 1991-2000 for female breast cancer and colorectal cancer. In addition male and female colorectal cancer were analysed for the twenty year period 1981-2000.

A matrix was calculated for all neighbouring MSOA for all 413 MSOA in Wales to enable the identity of all the neighbours. Table 4.7 shows the number of neighbours for all MSOAs in Wales.

No. of neighbours	No. of MSOA	No. of neighbours	No. of MSOA
1	5	7	59
2	23	8	22
3	45	9	14
4	66	10	9
5	77	11	2
6	90	12	1

Table 4.7: Number of neighbours for MSOAs in Wales.

Nearly 75% of MSOAs have 6 or less nearest neighbours with a mean of 5.4 neighbours. One MSOA had 12 neighbours. The larger the number of neighbours that a MSOA has, the more 'strength' the MSOA is able to borrow from neighbouring areas to smooth the relative survival rates. i.e. a smoothed MSOA with 1 neighbour has far less 'stability' over a smoothed MSOA with 6 neighbours.

The Bayesian smoothed survival rates in figure 4.7 show female breast cancer in Wales for the three time periods 1981-2000, 1981-1990 and 1991-2000 using the same equal intervals for all three maps.

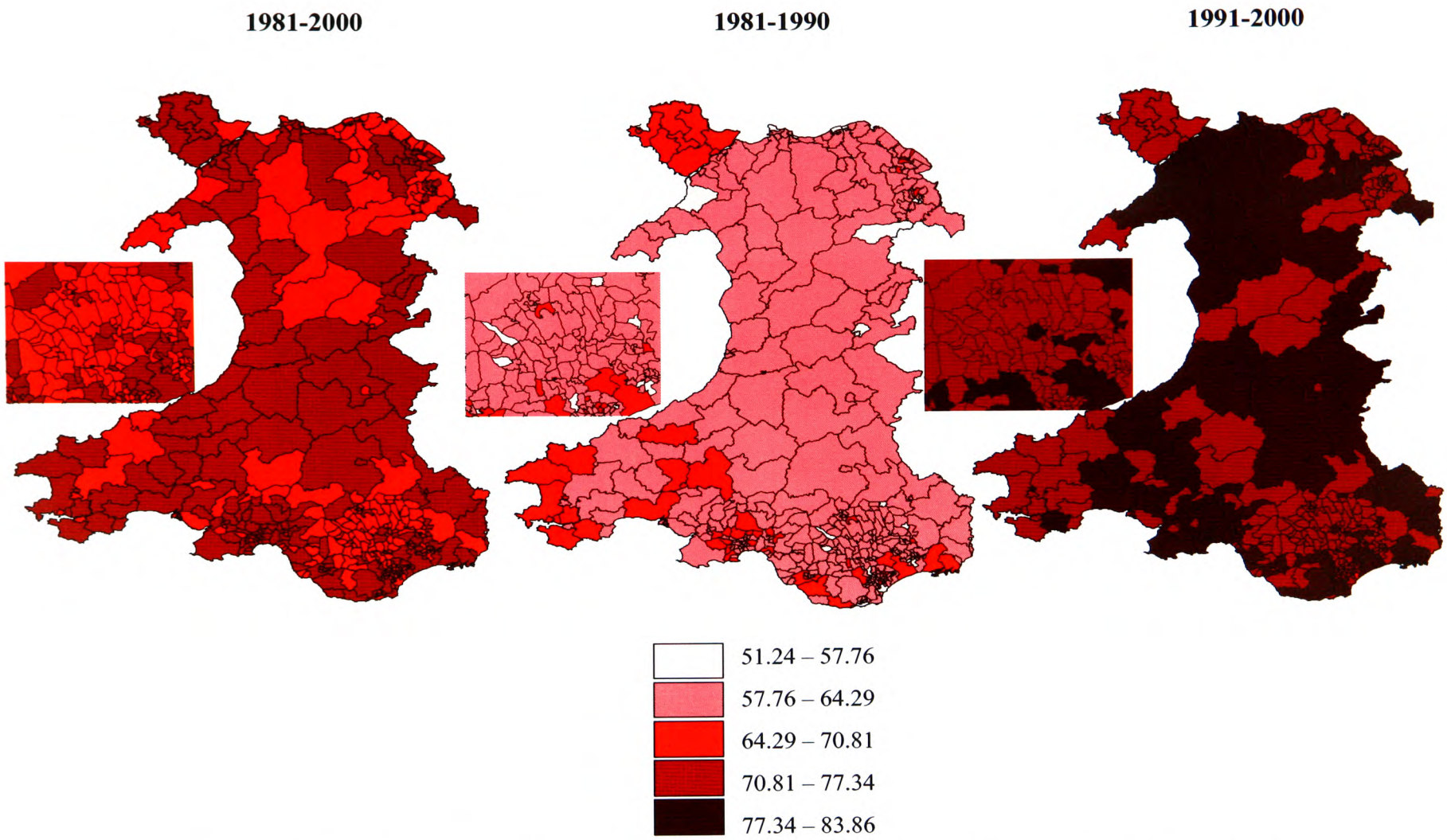


Figure 4.7: Bayesian smoothed five year relative survival for female breast cancer by MSOA in Wales.

The ranges of the relative survival rates in figure 4.7 for each of the periods are smaller compared with pre-smoothing. The ranges are 66.35% to 74.30% for the period 1981-2000, 51.24% to 69.28% for the period 1981-1990 and 68.58% to 83.86% for the period 1991-2000. For the period 1981-2000, there were just two colours on the map, with slightly higher five year relative survival rates in Mid-Wales and along the border of Wales. The first ten year period shows higher survival in Anglesey and parts of the South and West Wales coast. The second ten-year period shows a similar pattern to the twenty year period although survival rates have improved for most of the MSOAs compared with the period 1981-2000.

Figure 4.8 shows the corresponding Bayesian smoothed five year relative survival rates of colorectal cancer for the periods 1981-2000, 1981-1990 and 1991-2000.

Smoothing the survival rates has caused the ranges of the interval to diminish with a range between 22.60% and 45.02% for the period 1981-1990, a range of 12.48% to 42.73% for 1981-1990 and a range of 25.21% to 49.34% for the period 1991-2000. The worst survival rates were found in the South Wales Valleys for the twenty year period 1981-2000, although it appeared that the gap had decreased for the later ten year period 1991-2000 since the South Wales Valleys survival rates compared well with other parts of Wales. The highest survival rates were observed in Anglesey for the period 1991-2000 and for a few MSOA in Mid-Wales and along the South Wales coast. It was clear that compared with the initial ten year period, relative survival rates have improved for all MSOA throughout Wales for the period 1991-2000.

Figure 4.9 shows the Bayesian smoothed five year survival rates of colorectal cancer for persons, males and females for the period 1981-2000 by MSOA in Wales.

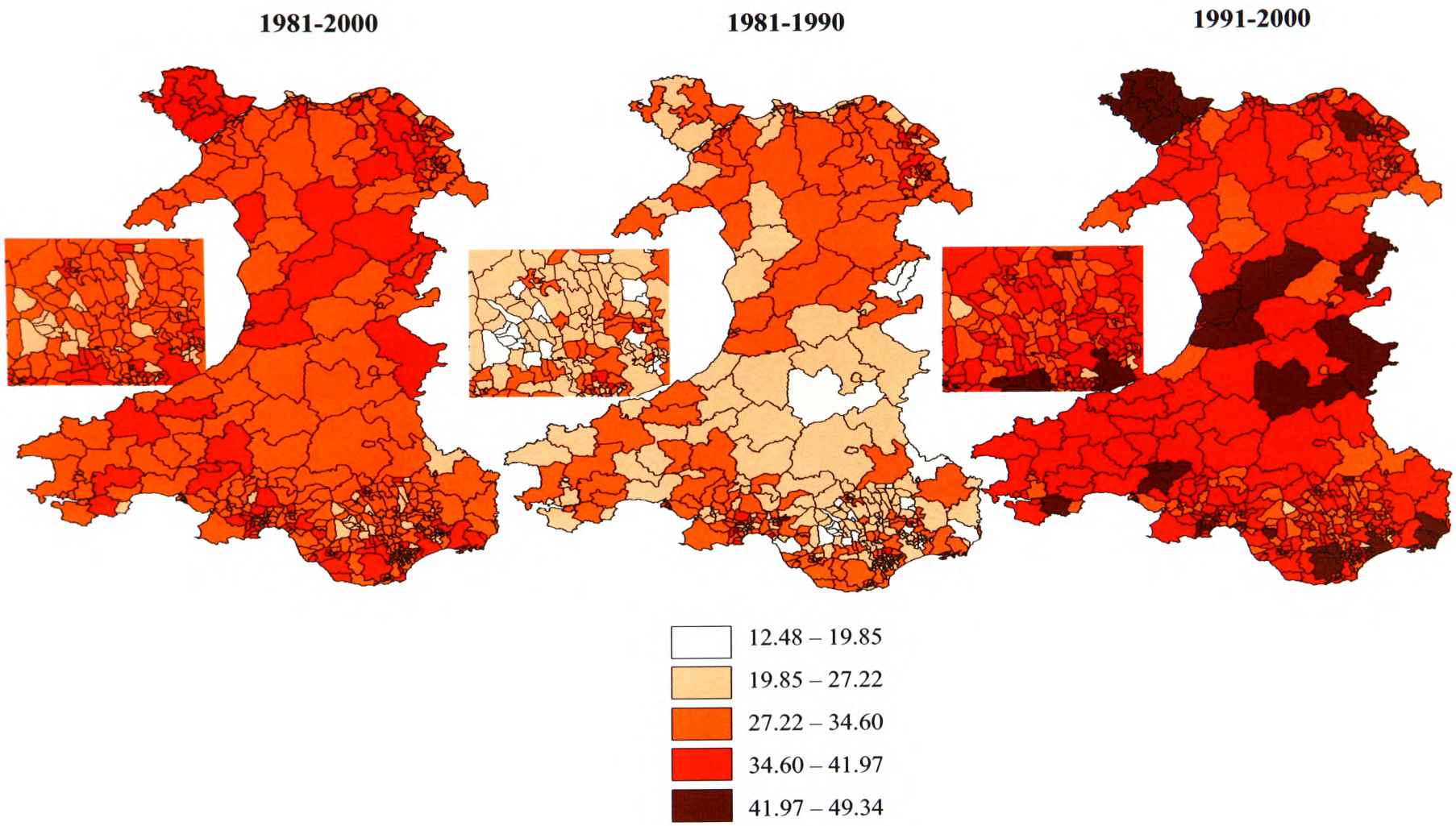


Figure 4.8: Bayesian smoothed relative survival by sex for colorectal cancer in Wales 1981-2000.

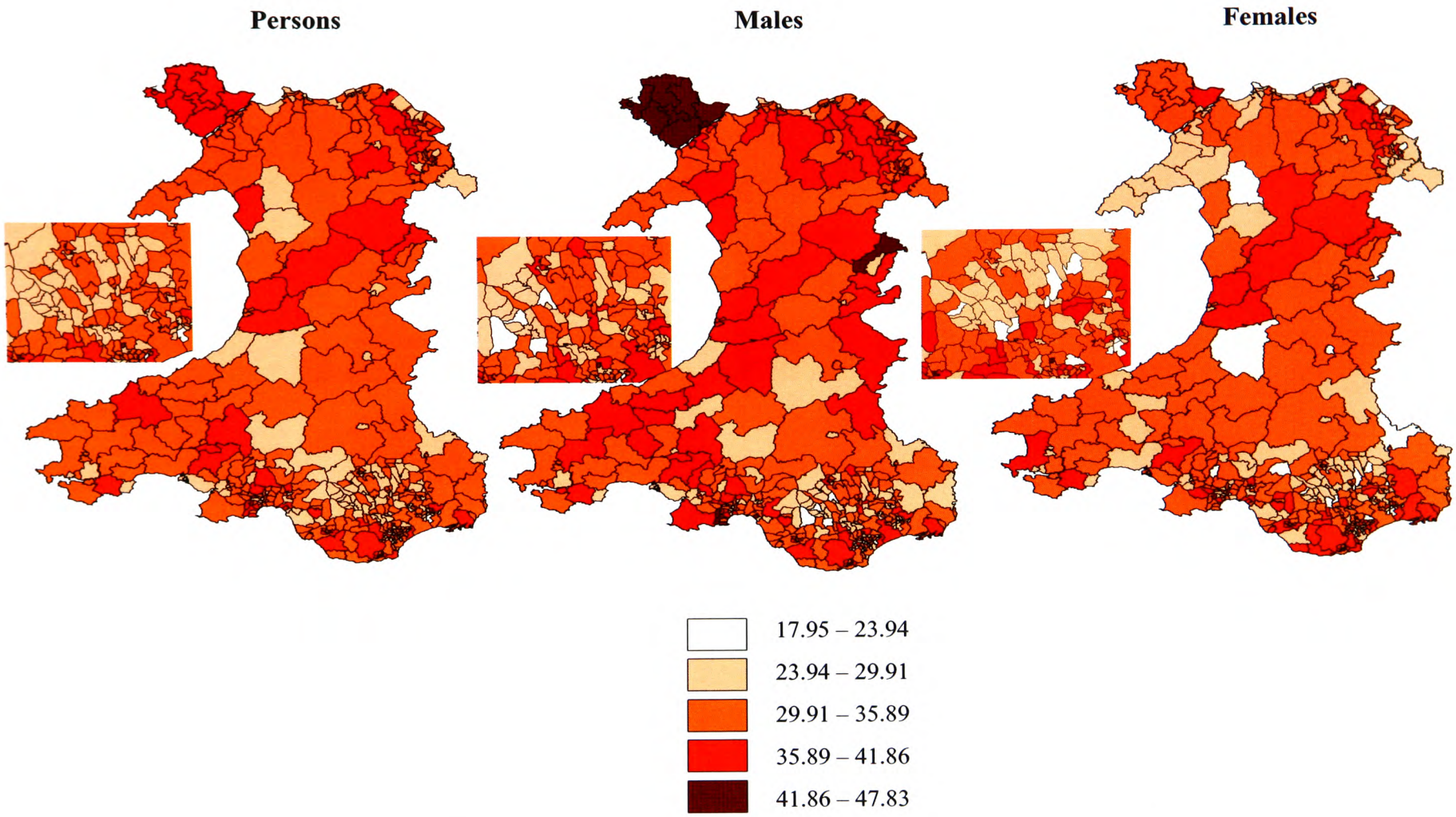


Figure 4.9: Bayesian smoothed relative survival by period for colorectal cancer in Wales 1981-2000.

There still does not appear to be any distinct pattern for colorectal cancer by sex, however, survival rates in Anglesey for males are higher than any other part of Wales and survival rates generally lower than most of Wales in the South Wales Rhondda Valleys.

Figure 4.10 shows the pre and post smoothed relative survival rates with the same break options for female breast cancer as opposed to all pre smoothing rates or post smoothing rates having the same break options.

Previously, the pre and post smoothed data compared the time periods to show how survival had increased throughout the periods examined by using the same scales. Here, the same scales for the pre and post smoothed data for each time period are used.

It can be seen for female breast cancer that before smoothing, the survival rates vary between neighbouring areas and it is difficult to determine where neighbouring areas of high and low survival rates exist. When the data are smoothed, nearly all MSOAs are contained in the same class for all time periods. There are a small number of "pockets" of higher and lower survival but in general it appears that the method has smoothed the data to the overall mean for Wales. This is explored in further detail in section 4.6.5.

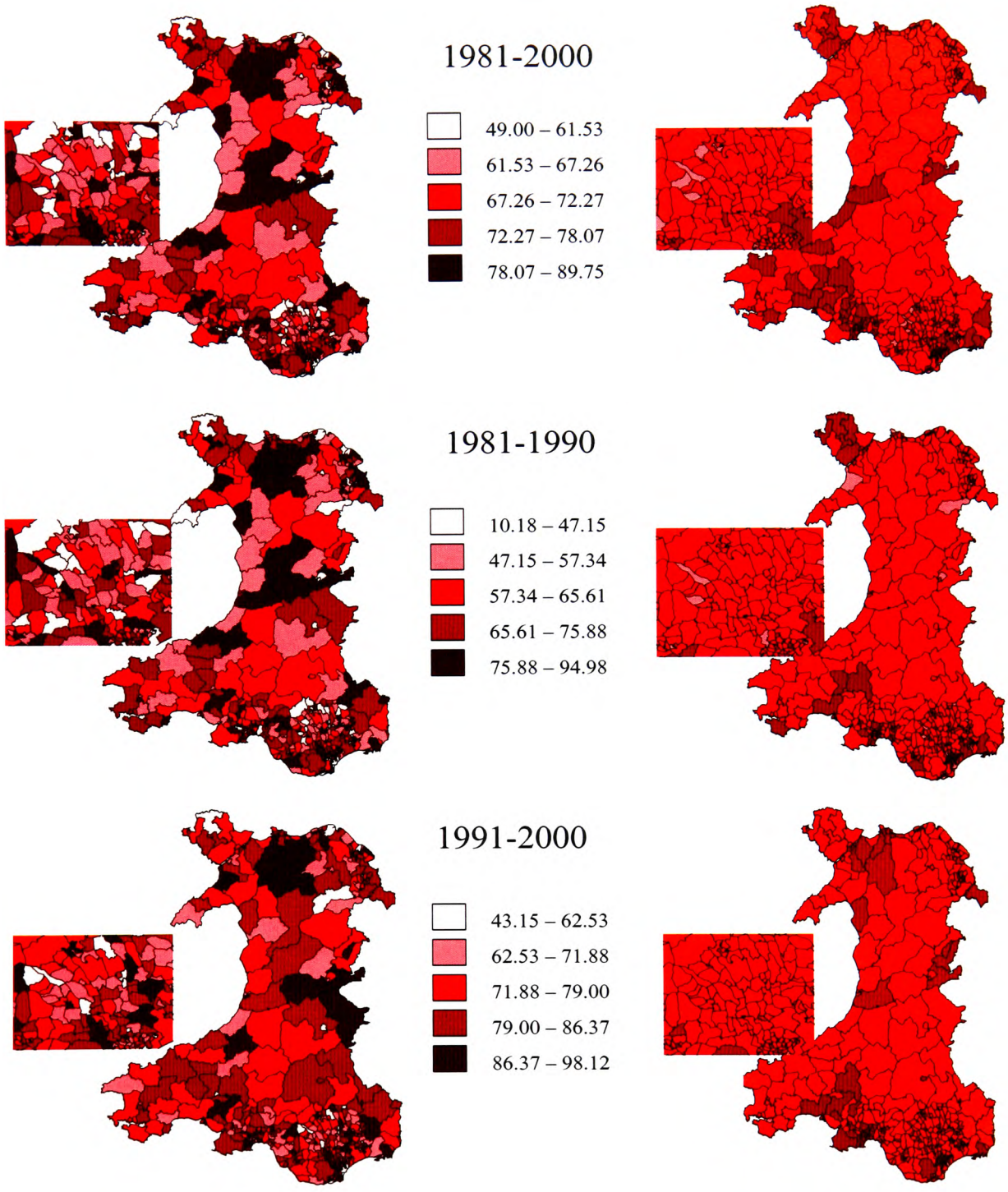


Figure 4.10: Pre and post smoothed survival rates using the same break options as original relative survival rates.

Figure 4.11 and figure 4.12 show the corresponding relative survival rates pre and post smoothing for colorectal cancer 1981-2000 in Wales for the three time periods and by sex.

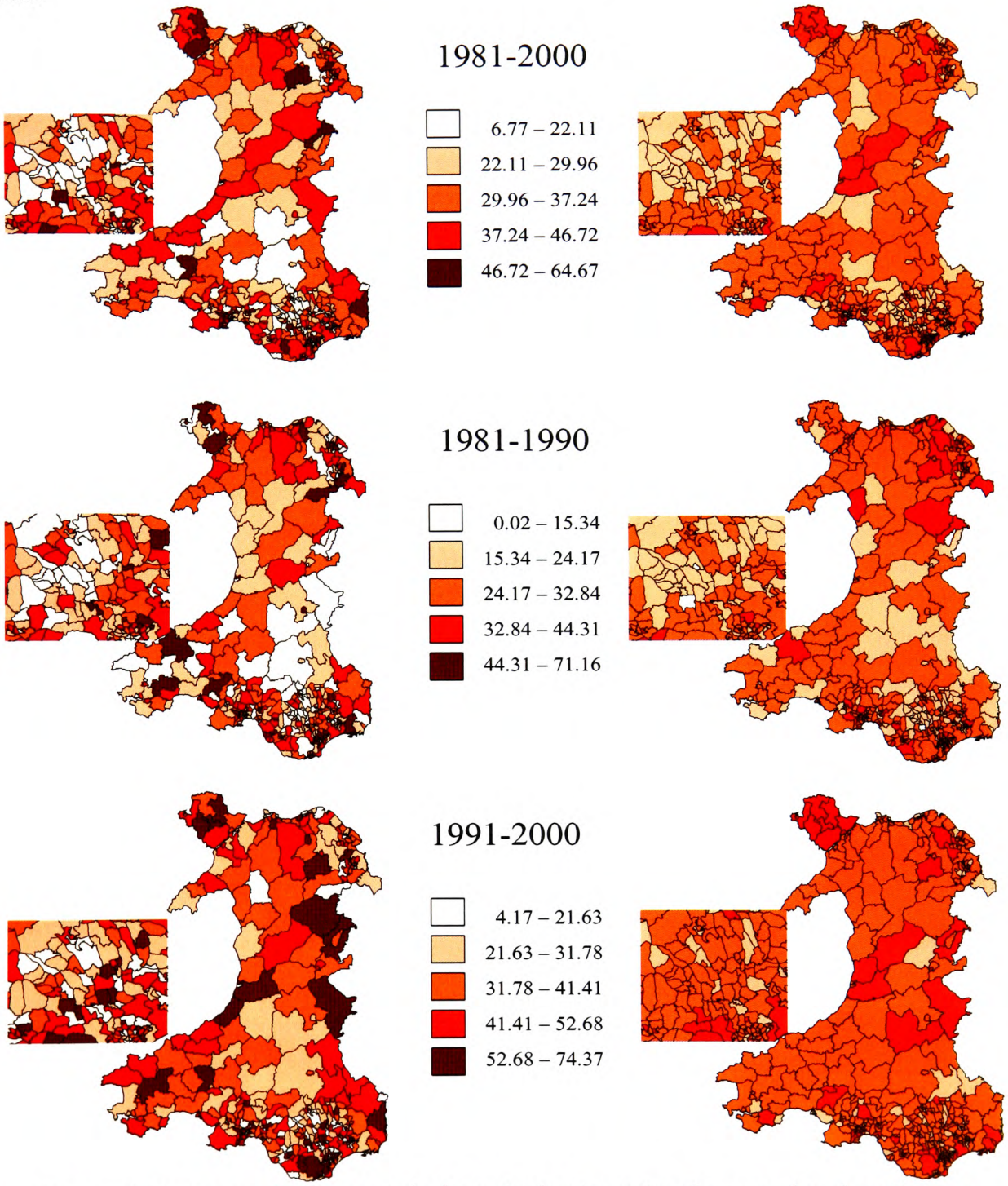


Figure 4.11: Pre and post smoothed survival rates using the same break options as original relative survival rates.

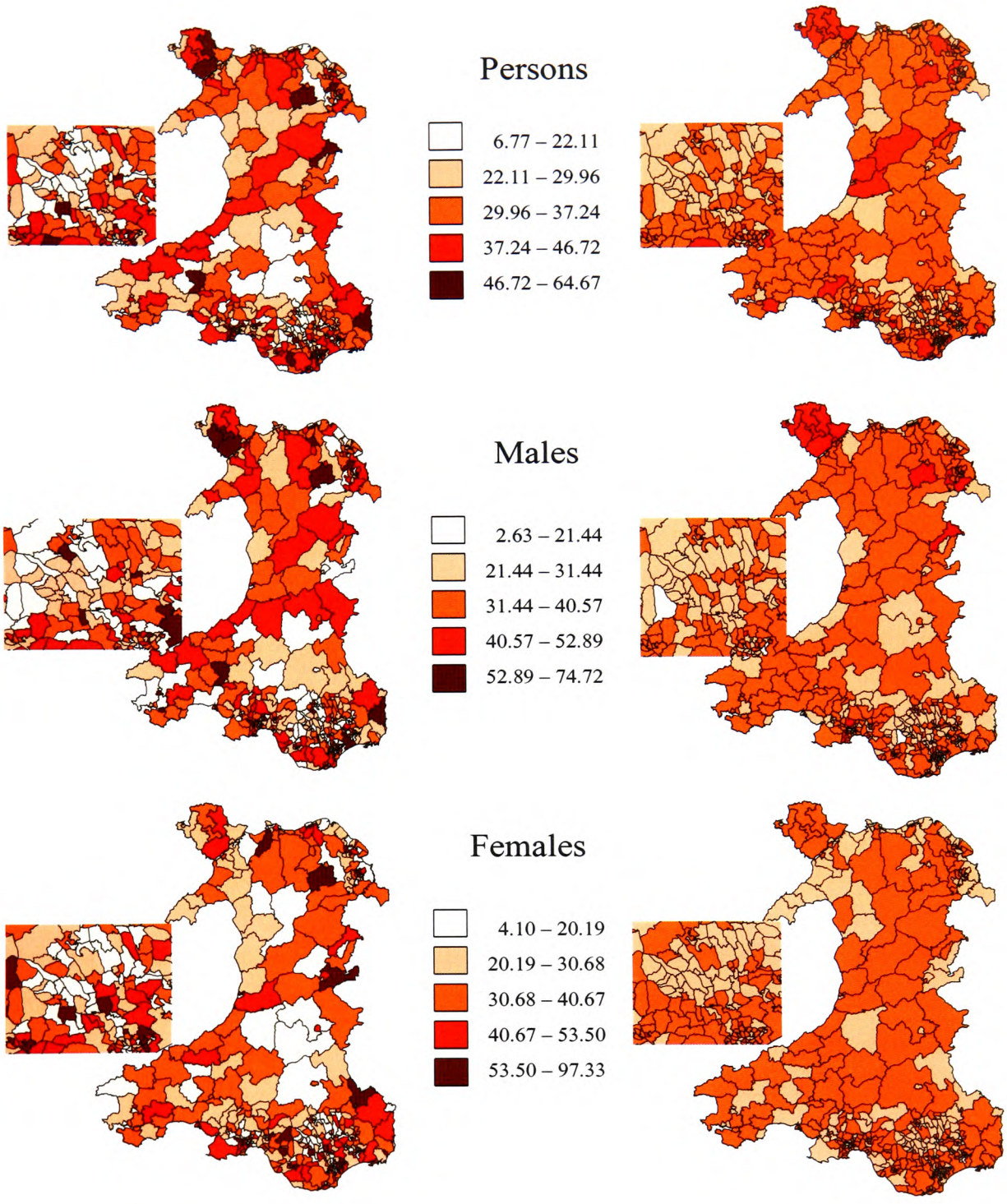


Figure 4.12: Pre and post smoothed survival rates using the same break options as original relative survival rates.

The same cannot be said for colorectal cancer as for female breast cancer regarding the similar relative survival rates for all geographical units. It is difficult to observe areas of high and low survival from the pre-smoothed data but from the smoothed data, it can be seen that lower survival rates were observed in the South Wales valleys for all periods and for both sexes. Also, higher survival rates were generally found in the northern areas of Wales and Anglesey.

Osnes et al found that Bayesian smoothing of female breast cancer resulted in localised areas of high and low survival and produced realistic clusters regarding the geographical detail. The results above show that female breast cancer smoothed close to the Wales average for all MSOAs – there is less smoothing towards the overall average for colorectal cancer, however it is difficult to observe localised areas of high and low survival since the maps are nearly all the same colour. For colorectal cancer, more localised clusters of high and low survival are observed. Additionally, there are more deaths for colorectal cancer compared with female breast cancer and so random variation is less likely to affect any real trends seen in the data.

In general, small area survival patterns for colorectal cancer appear more unstable compared with female breast cancer in terms of smoothing towards the overall average survival rate. i.e. localised clusters were found using colorectal cancer.

Moran's I method can be used to calculate the autocorrelation in the data to determine the extent to which neighbouring areas have similar or dissimilar rates – relative survival rates are used to investigate the autocorrelation. This may aid the interpretation of the above results. Table 4.8 shows the results of this analysis for female breast cancer and colorectal cancer in Wales for various time periods.

Cancer Site	Av dis freq	Moran's I	p-value
Female Breast 1981-2000	70.5125	0.0475	0.108
Female Breast 1981-1990	61.8460	0.0737	0.016
Female Breast 1991-2000	76.9023	0.0828	0.012
Colorectal 1981-2000	32.1004	0.0838	0.008
Colorectal (Males) 1981-2000	33.0776	0.0716	0.010
Colorectal (Females) 1981-2000	31.7628	0.0357	0.202
Colorectal 1981-1990	27.0615	0.0865	0.008
Colorectal 1991-2000	36.7604	0.0153	0.556

Table 4.8: Moran's I analysis of relative survival by MSOA in Wales.

All of Moran's I statistics are positive indicating that nearby areas have similar rates i.e. global spatial clustering is evident. The larger the value of Moran's I statistic the more similar that neighbouring rates are. Thus, the value of 0.0153 (non-significant) for colorectal cancer for the period 1991-2000 shows least similar rates for neighbouring areas whereas the value of 0.0838 (significant) for colorectal cancer for the period 1981-2000 shows the most similar rates for neighbouring areas. This is unusual, since from the previous smoothing analysis, female breast cancer smoothes closer to the overall mean (although the female breast cancer analysis for both ten year periods (both significant) are also very close to the high value for colorectal cancer). However, when smoothing colorectal cancer, areas of high and low clustering were found whereas for female breast cancer, as the survival rates were similar pre-smoothing, the smoothing did not show any areas of high and low clustering and tended to over smooth the data.

To summarise, the use of Moran's I statistic has complemented the results found in the earlier smoothing analysis. It should be noted that Moran's I statistic does not take into account the incidence and mortality figures by MSOA into the survival calculation like the survival analysis does – Moran's I just takes into account the survival rate itself.

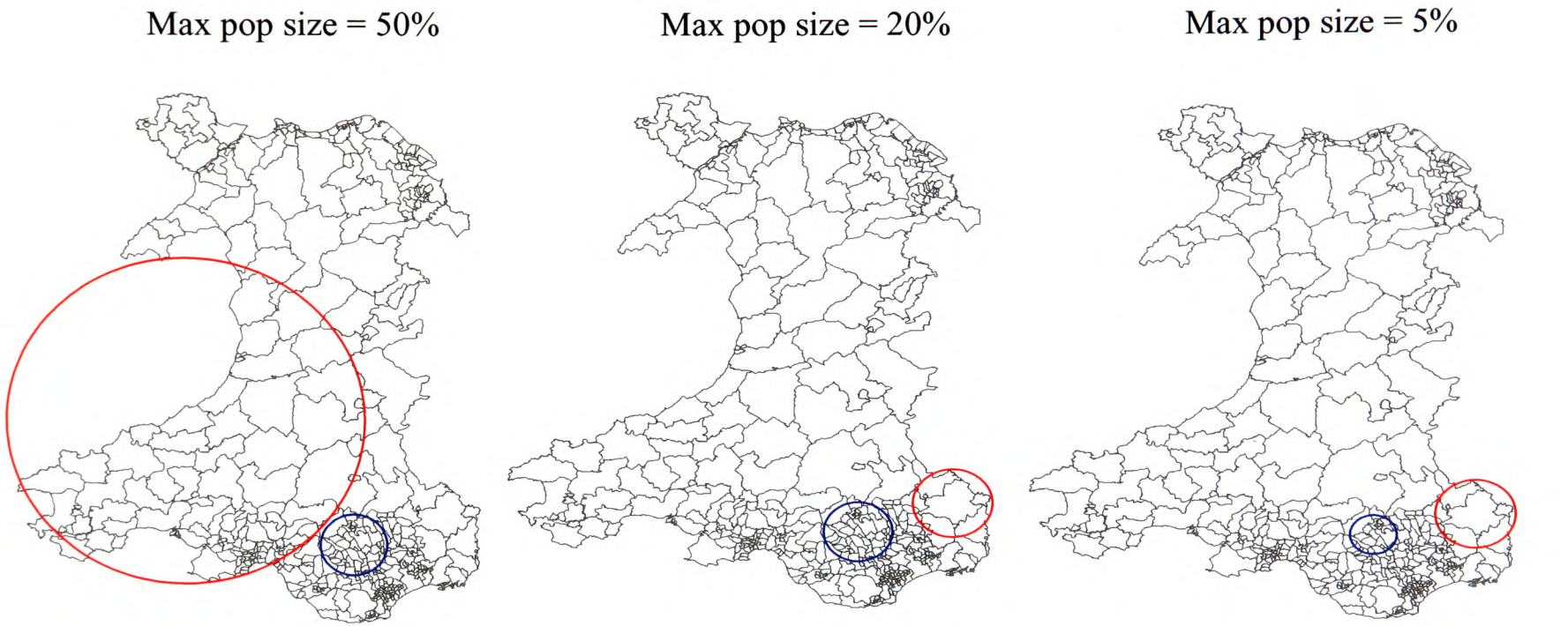
4.6.5. Local cluster analysis of survival

Theme one identified the spatial scan statistic by Kulldorff as the most effective algorithm to detect clusters. This method was used to determine whether the relative survival patterns found in the smoothed models were consistent with the high and low survival rates that the spatial scan statistic identified. The method is adjusted for the

Welsh cancer datasets to compare the survival rates in areas of Wales as opposed to the incidence in Wales. The relative survival rates for each MSOA were applied to the corresponding populations in each MSOA in order to obtain the number of cases that had survived for each MSOA to enter into the model as the cases. Thus, the number of cases that had survived from a "population" at risk of all cases was used for this analysis. The total number of cases represented the population at risk.

Figure 4.13 shows the results of the clustering algorithm when survival rates by MSOA were examined for the period 1981-2000. It is difficult to compare the results for female breast cancer due to the smoothing model showing the same classification of relative survival rates and hence no variation.

Figure 4.14 shows the analysis of the clustering algorithm using the colorectal cancer data for the period 1981-2000 showing general agreement between the spatial scan statistic results and the smoothed survival rates; higher survival rates in the area of Swansea and low survival rates in the South Wales valleys.



	Low	High
Number of MSOA	45	105
Radius	13.48km	71.05km
Rel surv in cluster	65.80%	72.15%
p-value	0.001	0.001

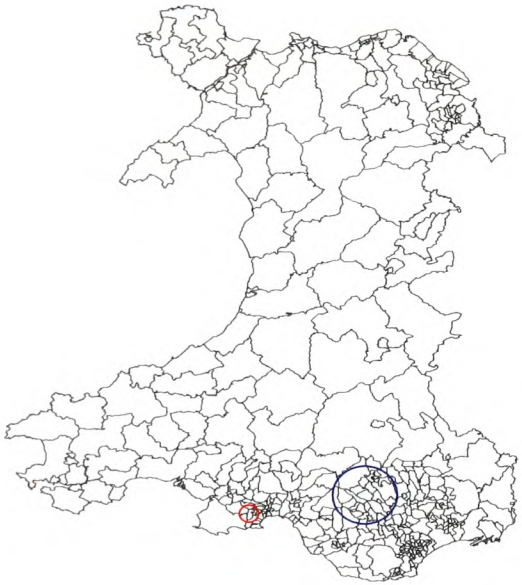
	Low	High
Number of MSOA	45	8
Radius	13.48km	15.43km
Rel surv in cluster	65.80%	78.01%
p-value	0.001	0.001

	Low	High
Number of MSOA	17	8
Radius	9.12km	15.43km
Rel surv in cluster	64.08%	78.01%
p-value	0.001	0.001

Low survival rate cluster High survival rate cluster

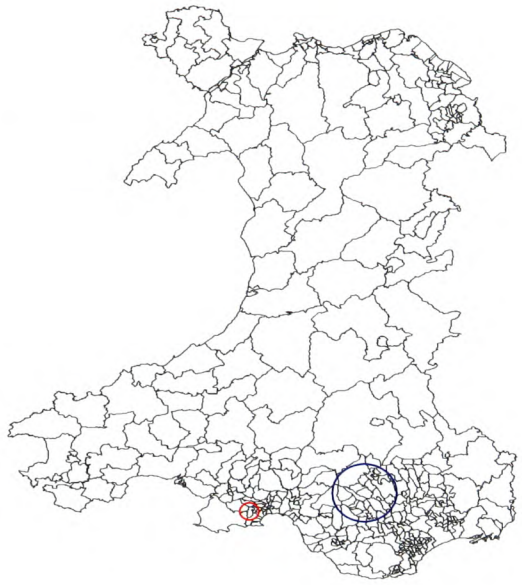
Figure 4.13: SaTScan analysis for female breast cancer in Wales 1981-2000.

Max pop size = 50%



	Low	High
Number of MSOA	29	6
Radius	11.94km	3.78km
Rel surv in cluster	22.48%	46.88%
p-value	0.001	0.001

Max pop size = 20%



	Low	High
Number of MSOA	29	6
Radius	11.94km	3.78km
Rel surv in cluster	22.48%	46.88%
p-value	0.001	0.001

Max pop size = 5%

	Low	High
Number of MSOA	7	6
Radius	1.54km	3.78km
Rel surv in cluster	14.73%	46.88%
p-value	0.001	0.001

Low survival rate cluster

High survival rate cluster

Figure 4.14: SaTScan analysis of colorectal cancer by MSOA in Wales 1981-2000.

Figure 4.15 shows the locations of all high and low clusters using the spatial scan statistic with a maximum population size for a cluster to be 50% of the total population at risk (50% was used to produce larger clusters to enable a comparison with the smoothed figures). There were 20 local clusters for female breast cancer and 14 local clusters for colorectal cancer but it is evident that female breast cancer appears to cluster in neighbouring regions compared with colorectal cancer. Over half (215 of 413 MSOAs – 52%) of all MSOAs were included as a low or high rate cluster for female breast cancer whereas 116 (28%) MSOAs were included as a low or high rate cluster for colorectal cancer. Comparing this with the smoothed survival rates they follow a similar pattern, clarifying the results obtained earlier. The clusters located for colorectal cancer were smaller than those located for female breast cancer – another reason why female breast cancer appears to display a smoother survival pattern compared with colorectal cancer, since although both have similar rates within neighbouring areas, female breast cancer rates are similar at a larger area compared to colorectal cancer. Thus, by identifying all clusters of female breast cancer and colorectal cancer in Wales using the spatial scan statistic, it can be seen that many of the high rate clusters for female breast cancer are in neighbouring areas, unlike for colorectal cancer. Thus when smoothing is applied, these areas will tend to smooth to very similar values and is the reason as to why the relative survival rates for female breast cancer are near to the overall mean in Wales. However, for colorectal cancer, the clusters found are scattered throughout Wales and could be the reason as to why the smoothing for colorectal cancer is not as "over smoothed" as female breast cancer. This pattern could not have been observed had it not been for the analysis at small area level.

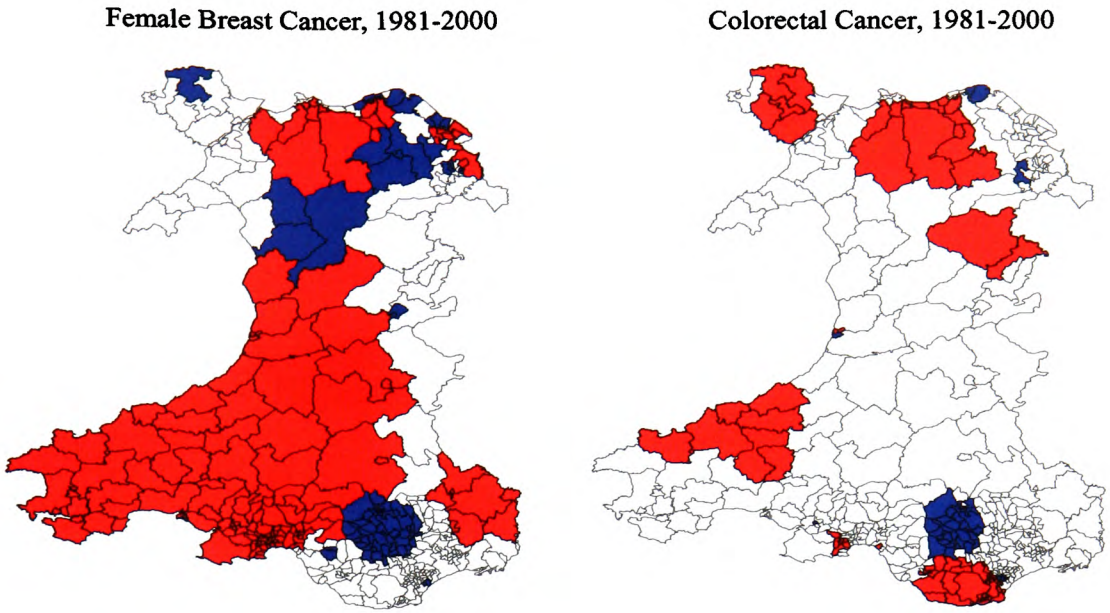


Figure 4.15: Locations of all high and low clusters using SaTScan for female breast cancer and colorectal cancer in Wales 1981-2000.

4.6.6. Breast cancer screening

Breast screening is a method of detecting breast cancer at a very early stage. The NHS Breast Screening Programme was set up to enable all women aged over 50 years in the UK to be provided with free breast screening every three years. In Wales, women were first screened in a few local health boards in 1989. Women can also be referred to a hospital breast clinic if aged under 50 years but this is not part of the NHS Breast Screening Programme. Table 4.9 shows the number of women that were screened for the period 1991-2000 in Wales. This data were obtained from Breast Test Wales.

	Number screened				% of women screened in	% of population in	% of population eligible for
Unitary Authority	Age <50	Age 50-64	Age 65+	Total	each UA	each UA screened	screening
Anglesey	500	14799	1023	16322	2.52	4.63	17.10
Gwynedd	928	22903	1921	25752	3.97	4.25	17.37
Conwy	773	22530	1659	24962	3.85	4.37	17.51
Denbighshire	615	18312	1447	20374	3.14	4.31	17.03
Flintshire	735	20292	1307	22334	3.45	3.02	16.12
Wrexham	896	22140	933	23969	3.70	3.70	16.29
Powys	684	30178	3183	34045	5.25	5.48	17.81
Ceredigion	619	14097	1358	16074	2.48	4.44	16.94
Pembrokeshire	894	25730	2486	29110	4.49	5.02	17.69
Carmarthenshire	1646	36437	2400	40483	6.25	4.57	18.06
Swansea	4864	45516	1888	52268	8.06	4.46	16.80
Neath Port Talbot	2592	26702	1180	30474	4.70	4.30	17.26
Bridgend	1131	32400	1656	35187	5.43	5.29	16.74
Vale of Glamorgan	1077	31803	1713	34593	5.34	5.70	16.37
Rhondda Cynon Taff	1752	46702	2972	51426	7.93	4.26	16.51
Merthyr Tydfil	664	10635	640	11939	1.84	3.96	16.45
Caerphilly	1485	32781	1843	36109	5.57	4.17	16.22
Blaenau Gwent	521	13911	1146	15578	2.40	4.22	16.61
Torfaen	1080	17740	1243	20063	3.10	4.28	16.73
Monmouthshire	725	18695	1402	20822	3.21	4.90	17.77
Newport	880	22065	1089	24034	3.71	3.43	16.49
Cardiff	1876	54496	2548	58920	9.09	3.75	14.88
UA not matched	77	2370	939	3386	0.52	NA	NA
Total	27014	583234	37976	648224	100.00	4.35	16.71

Table 4.9: Breast screening information from Breast Test Wales, 1991-2000.

Table 4.9 shows that the age group 50-64 contains the largest number of women who were screened since this is the target group. A total of 648,224 women were screened over the ten year period giving an average number of nearly 65,000 women per year being screened. The percentage of women aged between 50 and 64 (target group) in each Unitary Authority (UA – co-terminous with LHBs) that were actually screened varied from 1.84% in Merthyr Tydfil to 9.09% in Cardiff. However, when taking the female population into account for each LHB, the figures varied between 3.02% of women being screened in Flintshire to 5.70% of the total number of women in the Vale of Glamorgan being screened. The percentage of eligible women to be screened in each UA was input to the smoothing model so that each MSOA had its corresponding percentage of women of the total number of women in each UA. i.e. all MSOAs that were contained in Newport were assigned the figure 16.49.

Figure 4.16 shows the Bayesian smoothing of female breast cancer for the period 1991-2000 without the screening factor and with the screening factor.

The map on the left hand side in figure 4.16 shows the five year smoothed relative survival rates for the period 1991-2000 with the original natural breaks. The smoothed survival rates range from 68.58% to 83.86% by MSOA in Wales when screening was not taken into account. The central and right hand maps in figure 4.16 shows the corresponding five year smoothed relative survival rates by MSOA in Wales with the percentage of eligible women for screening for the period 1991-2000. These smoothed survival rates range from 72.83% to 79.90% in the new breaks, clearly an improvement in smoothing with the breast screening factor included. With the inclusion of the percentage of eligible women for screening into the smoothed model, there has been an improvement in survival in parts of West and South Wales with poorer survival in parts of North Wales.

A study by Anttinen et al concluded that detection by screening was not an independent prognostic factor regarding survival of breast cancer patients. It is clear that survival has improved in Wales over the two decades of analysis from 54% in 1981-1990 to 66% in 1991-2000 for five year relative survival of female breast cancer. This may be an indication of breast cancers being diagnosed earlier via screening and hence tumours being diagnosed at a much earlier stage than they would have previously.

4.7. Conclusions

To summarise, it was seen that relative survival rates for female breast cancer and colorectal cancer varied at LHB level in Wales. Relative survival rates for female breast cancer varied by as much as 10% between LHBs while for colorectal cancer, this figure was 18%. This warranted a closer inspection of relative survival rates at a lower level of geography. In the past, because of the numbers of deaths being very small, it was impracticable to calculate relative survival estimates at small area level. However, due to the introduction of Super Output Areas from the ONS, it has been possible, for the first

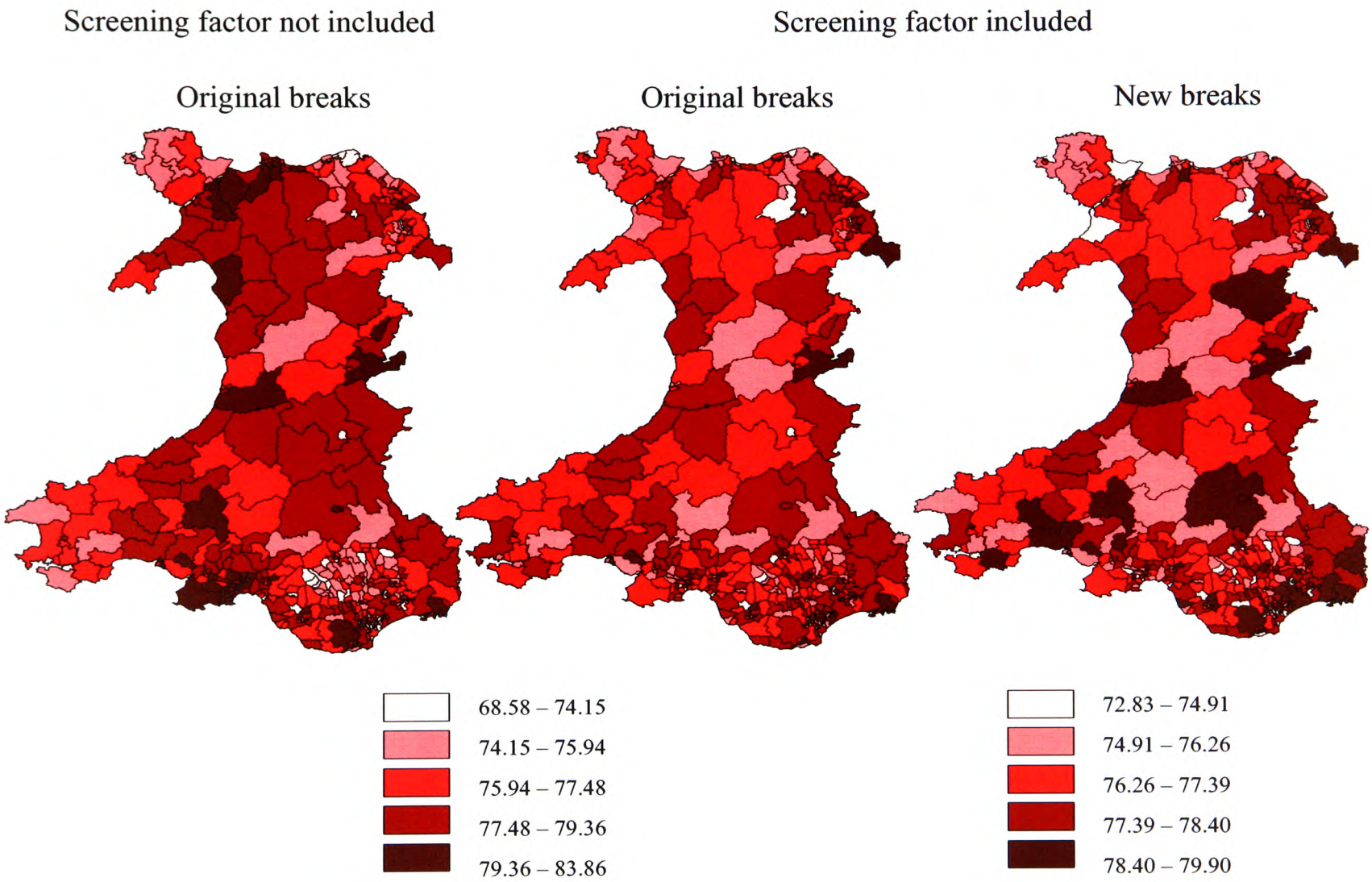


Figure 4.16: Bayesian smoothing of female breast cancer without and with breast screening factor, 1991-2000.

time, to calculate such estimates. Life tables were created to take into account age, sex and deprivation. One problem identified with calculating small area relative survival, was that the confidence intervals were large due to the small numbers. However, to overcome this issue, smoothing was examined. When smoothing female breast cancer relative survival rates, the maps showed one dominant colour, suggesting that the rates were "over smoothed". When smoothing the colorectal cancer survival rates, pockets of high and low areas of survival were found. When the spatial scan statistic was used to identify clusters, high survival rates were found in Monmouthshire and low survival rates were found in the South Wales Rhondda Valleys. For colorectal cancer, high survival rates were found in Swansea and low survival rates were located in Monmouthshire. The smoothed data also identified the South Wales Valleys as having lower survival rates compared with the rest of Wales for colorectal cancer. If all other clusters were located on a map of Wales using the spatial scan statistic, for female breast cancer, many high survival rate clusters were located near to each other and covered nearly half of Wales. This could be the reason as to why the smoothing for female breast cancer was over smoothed.

The percentage of women that were screened living in each UA varied between 1.8% and 9.1%. This screening factor was also included in the smoothing model for the period 1991-2000 to see if this had any effect on the smoothed relative survival rates. In general, the smoothed survival rates did not show any difference. If any change was observed, the figure was very small.

This theme has shown the importance of analysing data at small area level to identify areas of high and low survival that would not otherwise have been detected. Areas of future work in this theme include analysing distance from residence to treatment centre and using staging information (to name a few) which may explain the variation between these clusters of survival at small area level.

5. REFERENCES

Agency for Toxic Substances and Disease Registry (ATSDR) 2002. Public Health Investigations at the Nant-y-Gwyddon Landfill Gelli, Rhondda Cynon Taf, Wales: An Evaluation of the Environmental Health Assessment Process. Available at http://www.wales.nhs.uk/sites3/Documents/568/ATSDR%20final%20english.pdf Last Accessed June 19th 2008.

Ahlbom A, Day N, Feychting M , Roman E, Skinner J, Dockerty J, McBride M, Michaelis J, Olesen JH, Tynes T, Verkasalo PK. (2000) A pooled analysis of magnetic fields and childhood leukaemia. British Journal of Cancer, 83, 692–98.

Allgood PC, Bachmann MO (2006) Effects of specialisation on treatment and outcomes in screen-detected breast cancers in Wales: cohort study. British Journal of Cancer 94, 36-42.

Anselin L (1995) Local indicators of spatial association-LISA. Geographical Analysis 27: 93-115

Anselin L (2004) Review of cluster analysis software. In North American Association of Central Cancer Registries. Springfield, Illinois.

Anttinen J, Kautiainen H, Kuopio T (2006) Role of mammography screening as a predictor of survival in postmenopausal breast cancer patients. British Journal of Cancer 94, 147-151.

Bellec S, Hémon D, Rudant J, Goubin A, Clavel J (2006) Spatial and space-time clustering of childhood acute leukaemia in France from 1990 to 2000: a nationwide study. British Journal of Cancer, 94, 763-770.

Besag J, Newell J (1991) The detection of clusters in rare diseases. Journal of the Royal Statistical Society Series A 154: 143-155.

Bithell JF (1995). The choice of test for detecting raised disease risk near a point source. Statistics in Medicine 14: 2309-2322.

Bithell JF (1999) Disease mapping using the relative risk function estimated from areal data. Disease Mapping and Risk Assessment for Public Health. Lawson AB, Biggeri A, Bohning D, Lesaffre E, Viel JF, Bertollini R, eds. New York: John Wiley & Sons. pp. 247-55.

Black D (1984) Investigation of the possible increased incidence of cancer in West Cumbria: report of the Independent Advisory Group, HMSO, London.

Black RJ, Bashir SA. (1998) World standard cancer patient populations: a resource for comparative analysis of survival data. IARC Scientific Publications 145: 9-11.

Bland JM, Altman DG (1995) Multiple significance tests: the Bonferroni method. British Medical Journal. 310:170.

Boer JT, Pastor Jr M., Sadd JL, Synder LD (1997) Is there environmental racism? The demographics of hazardous waste in Los Angeles County, *Social Science Quarterly*, 78(4), 793-810.

Bonetti M, Pagano M (2001) On detecting clustering. Proceedings of the Biometrics Section, American Statistical Association, 24-33.

Bonetti M, Pagano M (2005) The interpoint distance distribution as a descriptor of point patterns, with an application to spatial disease clustering. Statistics in Medicine. 24:753–773.

Briggs D, de Hoogh K, Hurt C, Maitland I. (2001) A geographical analysis of populations living around landfill sites. London: Small Area Health Statistics Unit, Imperial College. (SAHSU Technical Report 2001.1.)

Briggs D, Fecht D, de Hoogh K (2007) Census data issues for epidemiology and health risk assessment: experiences from the Small Area Health Statistics Unit. Journal of the Royal Statistical Society 170, part 2, 355-378.

Campbell NC, Elliott AM, Sharp L, Ritchie LD, Cassidy J, Little J (2000) Rural factors and survival from cancer: analysis of Scottish cancer registrations. British Journal of Cancer 82:1863-1866.

Cartwright RA, Alexander FE, McKinney PA, Ricketts TJ (1990) Leukaemia and lymphoma. An atlas of distribution within areas of England and Wales 1984-1988. Leukaemia Research Fund, London.

Castresana J (2002) Genes on human chromosome 19 show extreme divergence from the mouse orthologs and a high GC content. Nucleic Acids Research 30: 1751-1756.

Ceccato V, Persson LO (2002). Dynamics of rural areas: an assessment of clusters of employment in Sweden. Journal of Rural Studies 18: 49-63.

Clayton D, Kaldor J (1987). Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. Biometrics, Sep. 43(3):671-81.

Coleman MP, Bell CMJ, Taylor HL, Primic-Zakelj M (1989) Leukaemia and residence near electricity transmission equipment: a case-control study. British Journal of Cancer 60: 793-798.

Coleman MP, Babb P, Sloggett A, Quinn M, De Stavola B (2001) Socioeconomic Inequalities in Cancer Survival in England and Wales. Statistics in Medicine 91: S1, 208-216.

Cullen LE, Stewart GH, Duncan RP, Palmer JG (2001). Disturbance and climate warming influences on New Zealand Nothofagus tree-line population dynamics. Journal of Ecology 89: 1061-1071.

Cuzick J, Edwards R (1990) Spatial clustering for inhomogeneous populations. Journal of the Royal Statistical Society Series B 52: 73-104.

Devine OJ, Halloran ME, Louis TA (1994) Empirical Bayes methods for stabilizing incidence rates prior to mapping. Epidemiology 5: 622–630.

Dickman PW, Gibberd RW, Hakulinen T (1997) Estimating potential savings in cancer deaths by eliminating regional and social class variation in cancer survival in the Nordic countries. Journal of Epidemiology and Community Health 51: 289-298.

Diggle PJ, Rowlingson BS (1994). A conditional approach to point process modelling of elevated risk. Journal of the Royal Statistical Society 157: 433-440.

Diggle PJ (1990). A point process modelling approach to raised incidence of a rare phenomenon in the vicinity of a prespecified point. Journal of the Royal Statistical Society 153: 349-362.

Dockerty JD, Sharples KJ, Borman B (1999). An assessment of spatial clustering of leukaemias and lymphomas among young people in New Zealand. Journal of Epidemiology and Community Health 53: 154-158.

Doll R, Peto R (1981) The causes of cancer. Journal National Cancer Institute; 66: 1191-1308.

Dolk H, Vrijheid M, Armstrong B, Abramsky L, Bianchi F, Garne E, et al (1998) Risk of congenital anomalies near hazardous-waste landfill sites in Europe: the EUROHAZCON study. Lancet 352: 423-427.

Draper G, Vincent T, Kroll ME, Swanson J (2005) Childhood cancer in relation to distance from high voltage power lines in England and Wales: a case-control study. British Medical Journal 330: 1290-1294.

Dunn CE, Kingham SP, Rowlingson B, Bhopal RS, Cockings S, Foy CJW, Acquilla SD, Halpin J, Diggle P, Walker D (2001) Analysing spatially referenced public health data: a comparison of three methodological approaches. Health and Place 7, 1, 1-12.

Dunn CE, Bhopal RS, Cockings S, Walker D, Rowlingson B, Diggle P (2007) Advancing insights into methods for studying environment–health relationships: A multidisciplinary approach to understanding Legionnaires' disease. Health and Place 13, 3, 677-690.

Elliott P, Briggs D, Morris S, de Hoogh C, Hurt C, Jensen TK, Maitland I, Richardson S, Wakefield J, Jarup L (2001) Risk of adverse birth outcomes in populations living near landfill sites. British Medical Journal 323:363-368.

Energy Network Associations (2007) Electric and magnetic fields: The facts. Last updated January 2007.
http://www.energynetworks.org/spring/SHE/pdfs/EMFs_070612.pdf
Last accessed June 19th 2008.

Estève J, Benhamou E, Croasdale M, Raymond L (1990). Relative survival and the estimation of net survival: elements for further discussion. Statistics in Medicine. May;9(5):529-38.

Farrow DC, Samet JM, Hunt WC (1996) Regional variation in survival following the diagnosis of cancer. Journal of Clinical Epidemiology 49: 843-847.

Feychting M, Ahlbom A. (1994) Magnetic fields, leukemia, and central nervous system tumors in Swedish adults residing near high-voltage power lines. Epidemiology 5:501-509.

Feychting M, Forssen U, Rutqvist LE, Ahlbom A (1998) Magnetic fields and breast cancer in Swedish adults residing near high voltage power lines. Epidemiology. 9(4): 392-397.

Fielder HMP, Poon-King CM, Palmer SR, Moss N, Coleman G (2000) Assessment of impact on health of residents living near the Nant-y-Gwyddon landfill site: retrospective analysis. British Medical Journal; 320: 19-22.

Fosgate GT, Carpenter TE, Chomel BB, Case JT, DeBess EE, Reilly KF (2002) Time-space clustering of human brucellosis, California, 1973-1992. Emerging Infectious Diseases 8: 672-678.

Gardner MJ, Snee MP, Hall AJ, Powell CA, Downes S, Terrell JD (1990) Results of case-control study of leukaemia and lymphoma among young people near Sellafield nuclear plant in West Cumbria. British Medical Journal, 300, 423-429.

Gilman EA, McNally RJQ, Cartwright RA (1999). Space-time clustering of acute lymphoblastic leukaemia in parts of the UK (1984-1993). European Journal of Cancer 35: 91-96.

Goldberg S, Siemiatyck J (1999) Risks of Developing Cancer Relative to living near a Municipal Solid Waste Landfill Site in Montreal, Quebec, Canada. Archives of Environmental Health. July/August 1999 Vol 54 (4): 291-296.

Grau HR (2002) Scale-dependent relationships between treefalls and species richness in a neotropical montane forest. Ecology 83: 2591-2601.

Gregorio DI, Kulldorff M, Sheehan TJ, Samociuk H (2004) Geographic distribution of prostate cancer incidence in the era of PSA testing. Urology 63, 78-82.

Hebert-Croteau N, Brisson J, Lemaire J, Latreille J, Pineault R (2005) Investigating the correlation between hospital of primary treatment and the survival of women with breast cancer. Cancer 104: 1343-1348.

Hjalmers U, Kulldorff M, Gustafsson G, Nagarwalla N (1996) Childhood leukaemia in Sweden: Using GIS and a spatial scan statistic for cluster detection. Statistics in Medicine Vol 15, 707-715.

Holland BS, MD Copenhaver (1987) An improved sequentially rejective Bonferroni test procedure. Biometrics 43: 417-23.

Huang L, Kulldorff M, Gregorio D (2007a) A Spatial Scan Statistic for Survival Data. Biometrics 63, 109-118.

Huang L, Pickle LW, Stinchcomb D, Feuer EJ (2007b) Detection of Spatial Clusters: Application to Cancer Survival as a Continuous Outcome. Epidemiology 18, 73-87.

ICF consulting (2002) Overview of the Potential for Undergrounding the Electricity Networks in Europe. Last updated February 2003.
http://ec.europa.eu/energy/electricity/publications/doc/underground_cables_ICF_feb_03.pdf
Last accessed June 19th 2008.

Jacquez GM (1994) User manual for Stat! Statistical software for the clustering of health events. Ann Arbor, MI: BioMedware.

Jacquez GM (1996) A k-nearest neighbor test for space-time interaction. Statistics in Medicine 15: 1935-49.

Jacquez GM, Greiling DA (2003) Local clustering in breast, lung and colorectal cancer in Long Island, New York. International Journal of Health Geographics 2:3.

Jacquez GM, Grieling DA (2003) Local clustering in breast, lung and colorectal cancer in Long Island, New York. International Journal of Health Geographies 2:3.

Jarup L, Briggs D, de Hoogh C, Morris S, Hurt C, Lewin A, Maitland I, Richardson S, Wakefield J, Elliott P (2002) Cancer risks in populations living near landfill sites in Great Britain. British Journal of Cancer 86(11): 1732-1736.

Jaworowski Z (1999) Radiation Risks and Ethics. Physics Today September 52 (9):24-29.

Jeffery JA, Ryan PA, Lyons SA, Thomas PT, Kay BH (2002). Spatial distribution of vectors of Ross River Virus and Barmah Forest virus on Russell Island, Moreton Bay, Queensland. Australian Journal of Entomology 41: 329-338.

Johnson GD (2004) Small area mapping of prostate cancer incidence in New York State (USA) using fully Bayesian hierarchical modeling. International Journal of Health Geographics. 3:29.

Juutilaninen J, Laara E, Pukkala E (1990) Incidence of leukaemia and brain tumours in Finnish workers exposed to ELF magnetic fields. International Archives of Occupational and Environmental Health. 62; 4, 289-293.

Kheifets L (2001) Electric and magnetic field exposure and brain cancer: A review. Bioelectromagnetics (Suppl 5):S120– S131.

Knox EG, Gilman EA (1997) Hazard proximities of childhood cancers in Great Britain from 1953-80, Journal of Epidemiology and Community Health 51: 151-159.

Knox EG (2000) Childhood cancers, birthplaces, incinerators and landfill sites. International Journal of Epidemiology 29: 391-397.

Kravdal O (1998) Social inequalities in cancer survival. Memorandum from Department of Economics, University of Oslo 3.

Kulldorff M (1997) A spatial scan statistic. Communs Statist. Theory Meth. 26 1481-1496.

Kulldorff M (2001) Prospective time periodic geographical disease surveillance using a scan statistic. J. R. Statistical Society A 164, Part 1, 61-72.

Kulldorff M, Mostashari F, Hartman J, Heffernan R (2005) A space-time permutation scan statistic for the early detection of disease outbreaks. PLoS Medicine, 2:216-224.

Kulldorff M, Tango T, Park P (2004) Power comparisons for disease clustering tests. Computational Statistics and Data Analysis 42: 665-684.

Kulldorff M, Huang L, Pickle L, Duczmal L (2006) An elliptic spatial scan statistic. Statistics in Medicine, 25(22):3929-3943.

Launoy G, Le Coutour X, Gignoux M, Pottier D, Dugleux G (1992) Influence of rural environment on diagnosis, treatment, and prognosis of colorectal cancer. Journal of Epidemiology and Community Health 46, 365-367.

Lawson AB (1989) Score tests for detection of spatial trend in morbidity data. Dundee: Dundee Institute of Technology.

Lawson AB, Biggeri A, Bohning D, Lesaffre E, Viel J-F, Bertollini R (1999) – Disease mapping and risk assessment for public health. Chichester, UK: Wiley.

Lawson AB (2001) Statistical Methods in Spatial Epidemiology. Chichester, UK: Wiley.

Li C-Y, Theriault G, Lin RS (1997) Residential exposure to 60-Hertz magnetic fields and adult cancers in Taiwan. Epidemiology 8 25-30.

Lipworth L, Abelin T, Connelly RR (1970) Socio-economic factors in the prognosis of cancer patients. Journal of Chronic Disorders 23, 105-115.

London S, Thomas D, Bowman JD, Sobel E, Cheng T-C, Peters J (1991) Exposure to residential electric and magnetic fields and risk of childhood leukaemia. American J Epidemiology 134: 923-937.

Machado-Coelho GLL, Assuncao R, Mayrink W, Caiaffa WT (1999). American cutaneous leishmaniasis in southeast Brazil: space-time clustering. International Journal of Epidemiology 28: 982-989.

Moran PAP (1950) Notes on continuous stochastic phenomena. Biometrika 37: 17-23

Mullee MA, De Stavola B, Romanengo M, Coleman MP (2004) Geographical variation in breast cancer survival rates for women diagnosed in England between 1992 and 1994. British Journal of Cancer 90, 2153-2156.

Murrin RJA, Harrison P, Neilson JR (2005) A highly unusual cluster of acute promyelocytic leukaemia: an environmental aetiology? Clinical and Laboratory Haematology 27:1 71-73.

New York State Department of Health (1998) Study of cancer incidences Surrounding Municipal Solid Waste Landfills, New York State. PB98-142144.

Norstrom M, Pfeiffer DU, Jarp J (2000) A space-time cluster investigation of an outbreak of acute respiratory disease in Norwegian cattle herds. Preventative Veterinary Medicine 47: 107-119.

Oden N (1995) Adjusting Moran's I for population density. Statistics in Medicine 14: 17-26.

O'Malley C, Le GM, Glaser S, Shema S, West DW (2003) Socioeconomic status and breast carcinoma survival in four racial/ethnic groups: A population study. Cancer 97, 1303-1311.

Openshaw, S., A. Turner, I. Turton, J. Macgill, and C. Brunsdon (2000) "Testing space-time and more complex hyperspace geographical analysis tools." Innovations in GIS 7: 87-100. London: Taylor & Francis.

Osnes K, Aalen OO (1999) Spatial smoothing of Cancer Survival: A Bayesian Approach. Statistics in Medicine 18: 2087-2099.

Palmer SR, Dunstan FDJ, Fielder H, Fone D, Higgs G, Senior M (2005) Risk of congenital Anomalies after the opening of Landfill Sites. Environmental Health Perspectives 113: 10, 1362-1365.

Pascutto C, Wakefield JC, Best NG, Richardson S, Bernardinelli L, Staines A, Elliott P (2000) Statistical issues in the analysis of disease mapping data. Statistics in Medicine 19: 2493-2519.

Pastor, M., Jr, Morello-Frosch, R. and Sadd, J.L. (2005) The air is always cleaner on the other side: Race, space, and ambient air toxics exposures in California, *Journal Of Urban Affairs,* 27(2), 127-148.

Perera FP (1977) Environment and Cancer: Who Are Susceptible?" Science 278, 5340: 1068-1073.

Preston-Martin S, Navidi W, Thomas D, Lee PJ, Bowman J, Pogoda J. (1996) Los Angeles study of residential magnetic fields and childhood brain tumors. American Journal of Epidemiology 143: 105-119.

Redfearn A, Roberts D (2002) Health Effects and Landfill Sites. Environmental and Health Impact of Solid Waste Management Activities. Issues in Environmental Science and Technology 18, Eds: R. E. Hester and R. M. Harrison. Cambridge, Royal Society of Chemistry 103-140.

Salvati M, Frati A, Russo N (2003) Radiation-induced gliomas: report of 10 cases and review of the literature. Surg Neurol 60:60-67.

Schmucki R, DeBlois S, Bouchard A, Domon G (2002). Spatial and temporal dynamics of hedgerows in three agricultural landscapes of southern Quebec, Canada. Environmental Management 30: 651-664.

Shapiro S, Coleman EA, Broeders M, Cood M, deKoonig H, Frachboud J, Moss S, Paci E, Stachenko E, Ballard-Barbash R (1998) Breast cancer screening programmes in 22 countries: current policies, administration and guidelines. International journal of Epidemiology 27: 735-742.

Sidak Z (1967) Rectangular confidence regions for the means of multivariate normal distributions. Journal of the American Statistical Association, 62, 626-633.

Simes RJ (1986) An improved Bonferroni procedure for multiple tests of significance. Biometrika. 73: 751-754.

Singh GK, Miller BA, Hankey BF, Edwards BK (2004) Persistent area socioeconomic disparities in US incidence of cervical cancer, mortality, stage and survival 1975-2000. Cancer 101, 1051-1057.

Sloggett A, Hills M, de Stavola B, Mander A (1999) strel - Estimation of relative survival [Stata program]

Song C, Kulldorff M (2003) Power evaluation of disease clustering tests. International Journal of Health Geographics 2:9.

Southern Medical Services Ltd, 2004.
http://www.southernmedicalservices.com.au/?page=sow.php&sq=YT1kYSZjaWQ9NTQ5JmFpZD05NTc5Jmg9aCZjcmM9MTE5MDE2NTc3Ng
Last accessed June 19th 2008.

Swartz JB (1998) An entropy-based algorithm for detecting clusters of cases and controls and its comparison with a method using nearest neighbours. Health and Place 4:67-77.

Tabar L, Yen MF, Vitak B, Chen H-H, Smith RT, Duffy SW (2003) Mammography service screening and mortality in breast cancer patients: 20 years of follow-up before and after the introduction of screening. Lancet 361: 1405-1410.

Tango T (1995) A class of tests for detecting 'general' and 'focused' clustering of rare diseases. Statistics in Medicine, 14, 2323-2334.

Townsend P, Phillimore P, Beattie A (1988) Health and Deprivation: Inequality and the North. London: Croom Helm.

Trent Cancer Registry (2005) How many deaths are needed for a reasonable survival estimate? Presented at UKACR conference.
http://www.empho.org.uk/Download/Public/10219/1/ukacr05_pbs_survival.pdf

Last accessed 12th March 2008.

Turnbull BW, Iwano EJ, Burnett WS, Howe HL, Clark LC (1990) Monitoring for clusters of disease: Application to leukemia incidence in upstate New York. American Journal of Epidemiology 132: S136-S143.

Twelves CJ, Thomson CS, Dewar JA, Brewster DH (2001) Variation in survival of women with breast cancer: Health Board remains a factor at 10 years. British Journal of Cancer 85: 637-640.

UK Childhood Cancer Study Investigators (1999) Exposure to power-frequency magnetic fields and the risk of childhood cancer. Lancet 354 (9194): 1925-1931.

UK Childhood Cancer Study Investigators (2000). The UK Childhood Cancer Study: objectives, materials and methods. British Journal of Cancer 82(5): 1073-1102.

UK Childhood Cancer Study Investigators (2002) Exposure to power frequency electric fields and the risk of childhood cancer in the UK. British Journal of Cancer 87(11): 1257-1266.

Van Buuren S, Zaadstra BM, Zwanikken CP, Buljevac D, van Noort JM (1998). Space-time clustering of multiple sclerosis cases around birth. Acta Neurologica Scandinavica 97: 351-358.

Verkasalo PK, Pukkala E, Hongisto M, Valjus J, Jarvinen P, Heikkila K, Koskenvuo M (1993) Risk of cancer in Finnish children living close to power lines. British Medical Journal 307: 895-899.

Verkasalo PK, Pukkala E, Kaprio J, Heikkila KV, Koskenvuo M. (1996) Magnetic fields of high voltage power lines and risk of cancer in Finnish adults: nationwide cohort study. British Medical Journal 313: 1047-1051.

von Winterfeldt D, Eppel T, Adams J, Neutra R, DelPizzo V(2004) Managing Potential Health Risks from Electric Powerlines: A Decision Analysis Caught in Controversy. Risk Analysis 24(6), 1487-1502.

Vrijheid M (2000) Health Effects of Residence Near Hazardous Waste Landfill Sites: A Review of Epidemiologic Literature. Environmental Health Perspectives 108(S1): 101-112.

Vrijheid M, Dolk H, Armstrong B, et al. (2002) Risk of chromosomal congenital anomalies in relation to residence near hazardous waste landfill sites in Europe. Lancet; 359:320-2.

Waller LA, Turnbull BW, Clark LC, Nasca P. (1992) Chronic disease surveillance and testing of clustering of disease and exposure: Application to leukemia incidence and TCE-contaminated dumpsites in upstate New York. Environmetrics 3: 281-300.

Waller LA, and Turnbull BW (1994) The effect of scale on tests of disease clustering. Statistics in Medicine 12: 1969-1984.

Waller LA, Turnbull BW, Clark LC, Nasca P (1994) Spatial pattern analyses to detect rare disease clusters. In Case Studies in Biometry. N. Lange, L. Ryan, L. Billard, D. Brillinger, L. Conquest, and J. Greenhourse eds. New York: John Wiley & Sons. pp. 13-16.

Waller LA, Carlin BP, Xia H, Gelfand AE (1997) Hierarchical spatiotemporal mapping of disease rates. Journal of the American Statistical Association 92: 607–617.

Wartenburg D (2001) The potential impact of bias in studies of residential exposure to magnetic fields and childhood leukaemia. Bioelectromagnetics Supplement 5: S32-S47.

Welsh Cancer Intelligence and Surveillance Unit (2001) Steward JA, Wright M, White C, Wade R (2001) Public concerns regarding the effect of Nant-y-Gwyddon Landfill Site (NYG) on the incidence of Non Hodgkin's Lymphoma (NHL) in the South Wales Rhondda Valley. Last updated March 2007.
http://www.wales.nhs.uk/sites3/Documents/242/NYG_report.pdf
Last accessed June 19th 2008.

Welsh Cancer Intelligence and Surveillance Unit (2002) Childhood cancer incidence in the Chepstow area 1974-1998.
http://www.wales.nhs.uk/sites/documents/242/PartI%2Epdf
Last accessed June 19th 2008.

Welsh Cancer Intelligence and Surveillance Unit (2002) Female breast cancer and prostate cancer in the Chepstow area 1985-1998.
http://www.wales.nhs.uk/sites/documents/242/PartII%2Epdf
Last accessed June 19th 2008.

Welsh Cancer Intelligence and Surveillance Unit (2002). Cancer Incidence, Mortality and Survival in Wales, 1993-2002.
http://www.wales.nhs.uk/sites3/page.cfm?orgid=242&pid=18138
Last accessed June 19th 2008.

Welsh Cancer Intelligence and Surveillance Unit (2004) Public concerns regarding the effect of Nant-y-Gwyddon Landfill Site (NYG) on the incidence of Non Hodgkin's Lymphoma (NHL) in the South Wales Rhondda Valley.
http://www.wales.nhs.uk/sites3/Documents/242/NYG%5Freport.pdf
Last accessed June 19th 2008.

Welsh Cancer Intelligence and Surveillance Unit (2005) Childhood leukaemia, brain tumours and retinoblastoma near the Menai Straits, North Wales 2000-2003.
http://www.wales.nhs.uk/sites/documents/242/050330MenaiReport.pdf

Last accessed June 19th 2008.

Wertheimer N, Leeper E (1979) Electrical wiring configurations and childhood cancer. Am J Epidemiol 109: 273-284.

Whittemore AS, Friend N, Brown BW, Holly EA (1987) A test to detect clusters of disease. Biometrika, 74:631-635.

Willett WC (1995) Diet, nutrition and avoidable cancer. Environment Health Perspective; 103(suppl 8): 165-170

Wrigley H, Roderick P, George S, Smith J, Mullee M, Goddard J (2003) Inequalities in survival from colorectal cancer: a comparison of the impact of deprivation, treatment, and host factors on observed and cause specific survival. Journal of Epidemiology and Community Health 57, 301-309.

Yin SN, Hayes RB, Linet MS, Li GL, Dosemeci M, Travis LB, Li CY, Zhang ZN, Li DG, Chow WH, Wacholder S, Wang YZ, Jiang ZL, Dai TR, Zhang WY, Chao XJ, Ye PZ, Kou QR, Zhang XC, Lin XF, Meng JF, Ding CY, Zho JS, Blot WJ (1996) A cohort study of cancer among benzene-exposed workers in China: overall results. Am J Ind Med. 29: 3 227-235

Youngson JHAM, Clayden AD, Myers A, Cartwright RA (1991) A case/control study of adult haematological malignancies in relation to overhead powerlines. British Journal of Cancer 63: 977-985.

Yu XQ, O'Connell DL, Gibberd RW, Smith DP, Dickman PW, Armstrong BK (2004) Estimating region variation in cancer survival: a tool for improving cancer care. Cancer Causes and Control 15: 611-618.

Yu XQ, O'Connell DL, Gibberd RW, Armstrong BK (2005) A population-based study from New South Wales, Australia 1996-2001: Area variation in survival from colorectal cancer. European Journal of Cancer (IN PRESS).

6. APPENDICES

6.1. Appendix A: Cluster Questionnaire

Four questions were sent to each cancer registry in the UKACR. The questions were as follows. Appendix A gives a summary of the responses.

1. When analysing alleged cancer clusters, does your cancer registry implement any clustering algorithms to identify a possible cluster? (e.g. the software SaTScan V5.1.3 is a spatial and space-time algorithm derived by Kulldorff. Other clustering algorithms include Knox, Local Moran, Besag and Newell, Turnbull to name a few)

2. If yes, what software and clustering algorithms does your cancer registry use to identify possible clusters?

3. If no, has your cancer registry ever considered using clustering algorithms to identify clusters?

4. Has your cancer registry ever conducted an analysis of various clustering algorithms to determine one that should be used for cluster analysis?

UKACR Cancer Registry	Q1	Q2	Q3	Q4
East Anglia Cancer Registry	No	-	No	No
Mersey and Cheshire Cancer Registry	No	-	Yes	No
Northern Ireland Cancer Registry	No	-	Yes	No
Northern & Yorkshire Cancer Registry and Information Service	No	-	No	No
North Western Cancer Registry	No	-	No	No
Oxford Cancer Intelligence Unit	No	-	No	No
Scotland Cancer Registry	No	-	Yes	No
South and West Cancer Intelligence Unit	No	-	No	No
Thames Cancer Registry	No	-	No	No
Trent Cancer Registry	No	-	No	No
West Midlands Cancer Intelligence Unit	No	-	No	No
Welsh Cancer Intelligence and Surveillance Unit	Yes	SaTScan	-	No

Appendix A – Analysis of UKACR Cluster Questionnaire.

Any other comments:

Mersey and Cheshire do not currently use any clustering algorithms but would like to. They are currently looking into using SaTScan V5.1.3 and Geo R.

Northern Ireland stated the possibility of purchasing SaTScan V5.1.3 was discussed and rejected some time ago.

Scotland - Bayesian smoothing techniques have been explored using Small Area Health Statistics Unit's (SAHSU) Rapid Inquiry Facility Software.

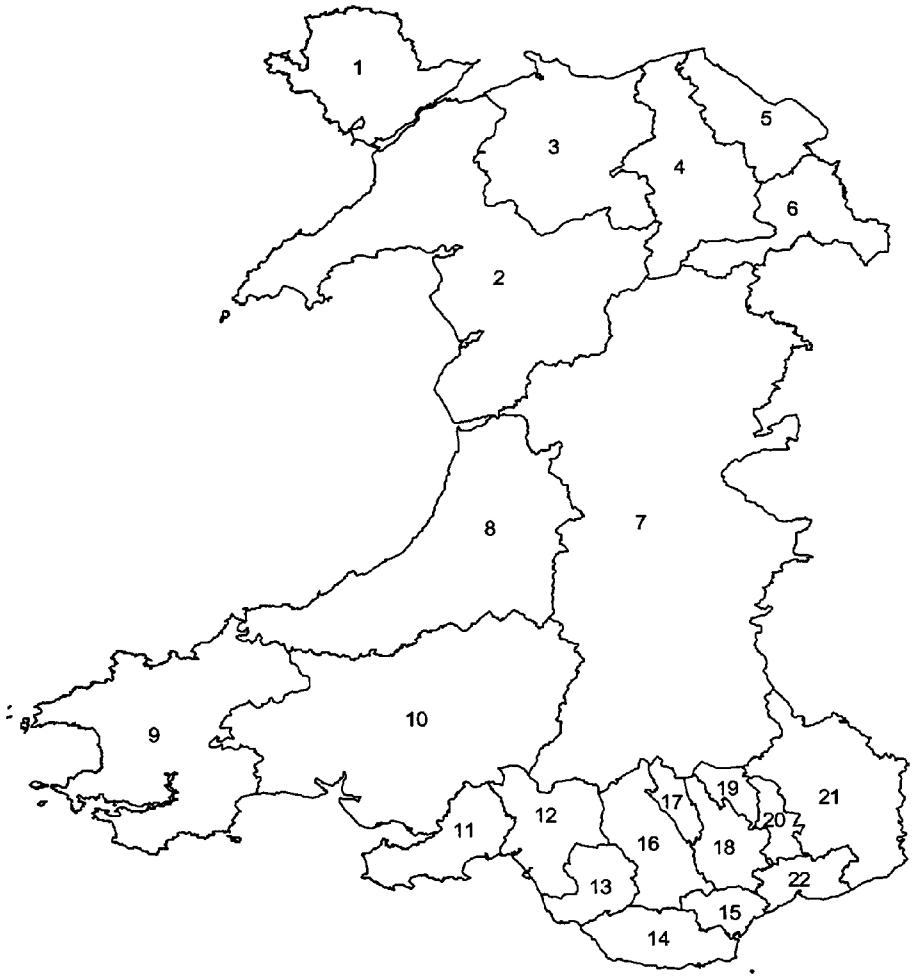
Thames tend to consult colleagues at the SAHSU at Imperial College on these matters.

Trent - These algorithms search for unidentified clusters (i.e. existence of clustering). Given that the cause of well established clusters is generally a mystery we regard looking for clusters as futile and a waste of time.

West Midlands have not had to use clustering algorithms as of yet. They work in collaboration with the Primary Care Trust involved. If complex analyses were required, they would probably consult with statisticians/ epidemiologists at the local university first.

Wales use SaTScan V5.1.3 to locate a cluster to determine if an alleged cancer cluster area is shown to be part of the most likely cluster and whether the area is significant or not.

6.2. Appendix B: Local Health Boards in Wales

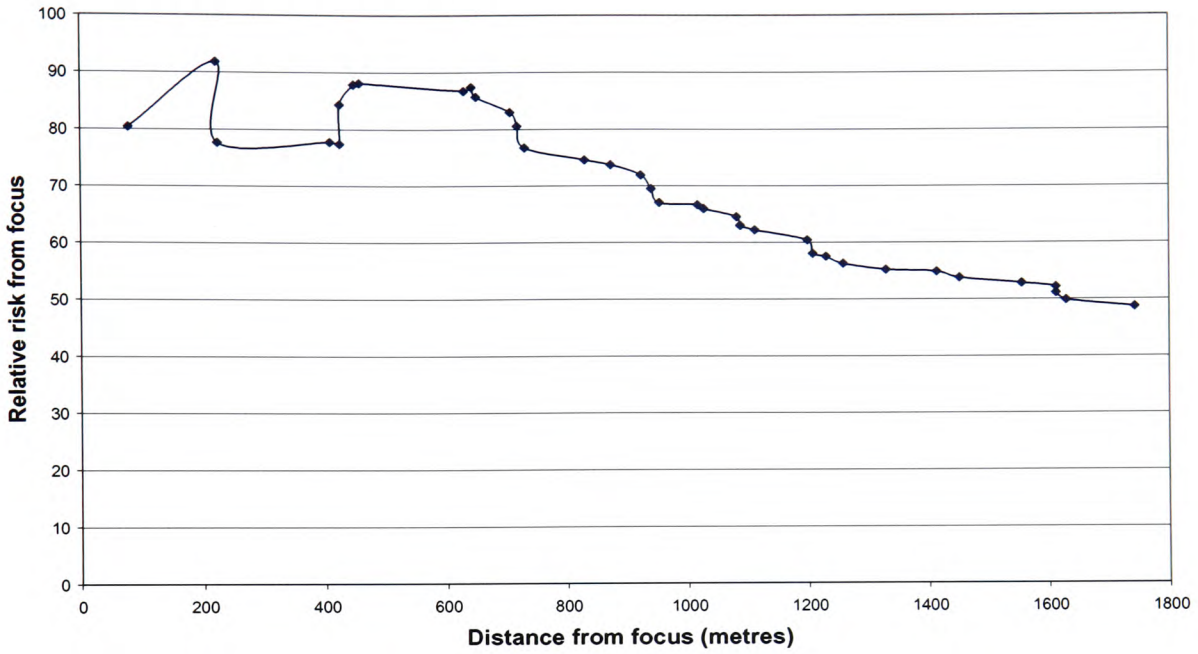


1	Isle of Anglesey	12	Neath & Port Talbot
2	Gwynedd	13	Bridgend
3	Conwy	14	Vale of Glamorgan
4	Denbighshire	15	Cardiff
5	Flintshire	16	Rhondda Cynon Taff
6	Wrexham	17	Merthyr Tydfil
7	Powys	18	Caerphilly
8	Ceredigion	19	Blaenau Gwent
9	Pembrokeshire	20	Torfaen
10	Carmarthenshire	21	Monmouthshire
11	Swansea	22	Newport

6.3. Appendix C: Relative risks of clusters within simulated datasets

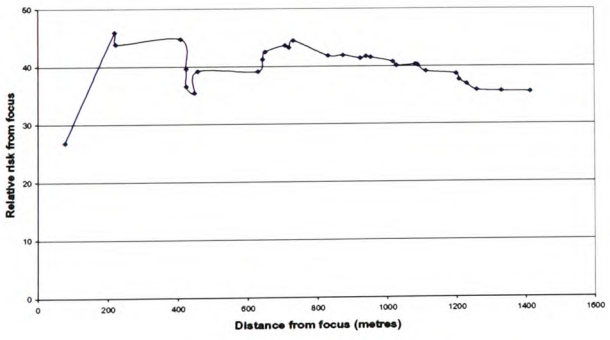
Stan Openshaw's datasets

Dataset 02: Focus 434500, 396000

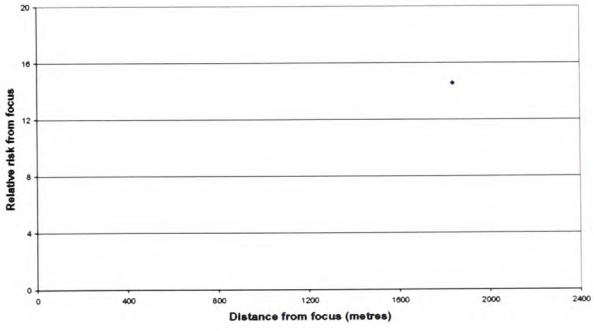


Dataset 04

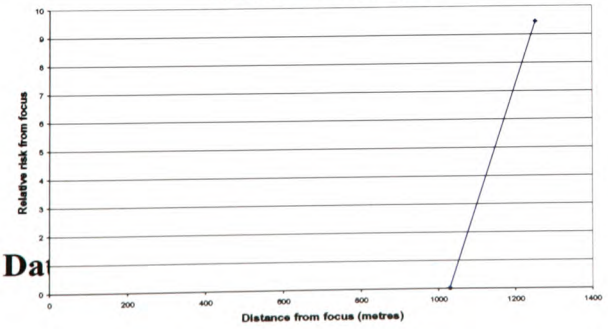
Focus: 434500, 396000



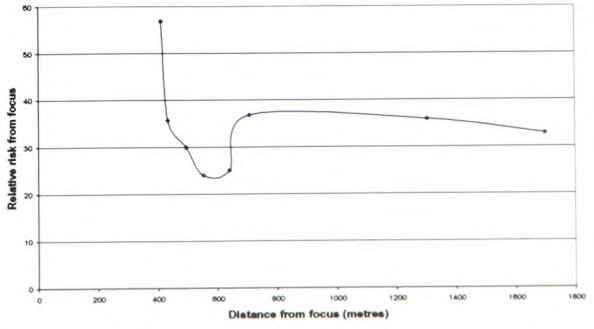
Focus: 483000, 472000



Focus: 447000, 446000

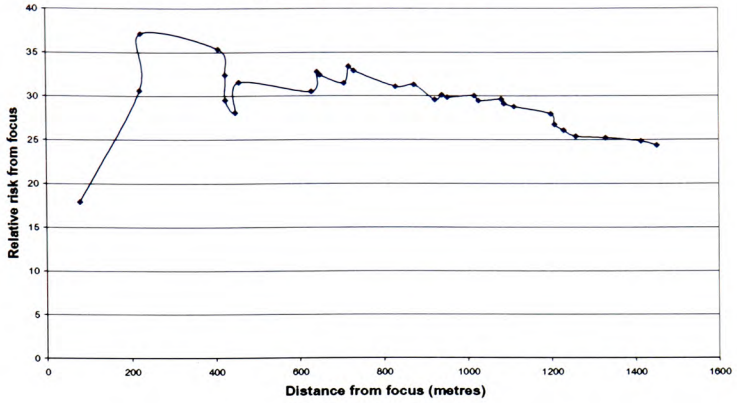


Focus: 459000, 457000

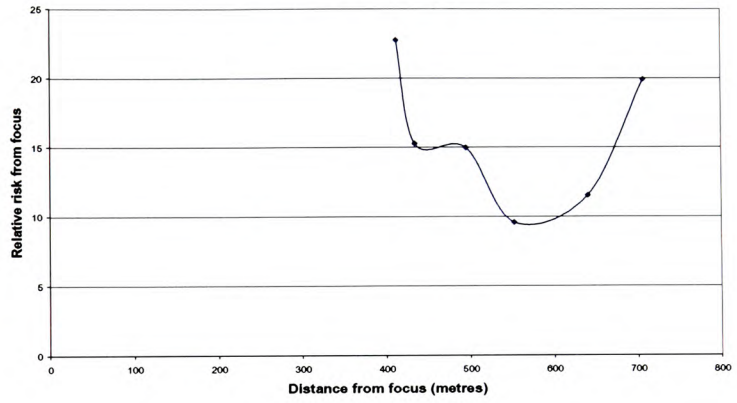


Da

Focus: 434500, 396000



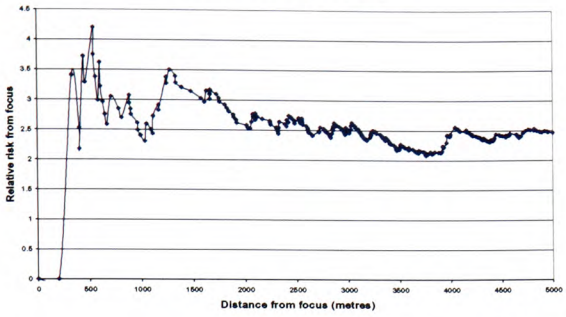
Focus: 459000, 447000



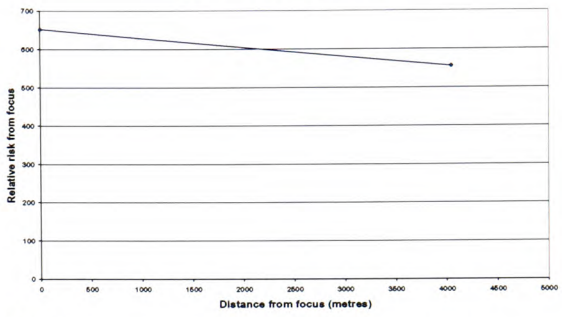
Chris Brunsdon's datasets

Dataset 1

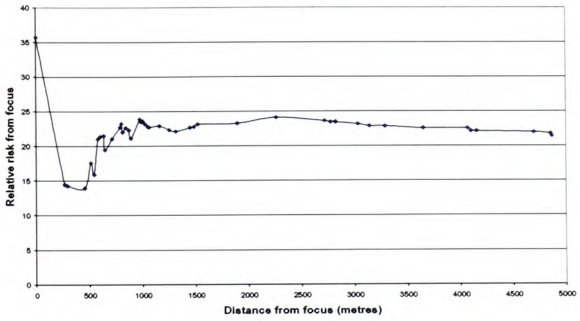
Focus: 414500, 396000



Focus: 461800, 495700

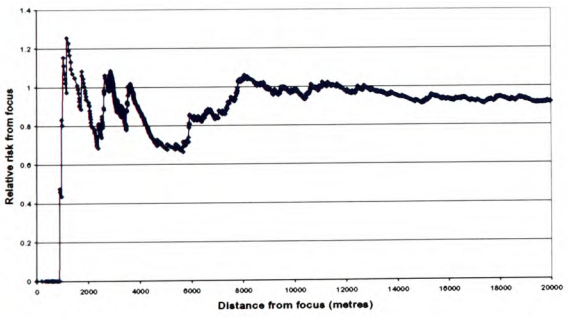


Focus: 436950, 493540

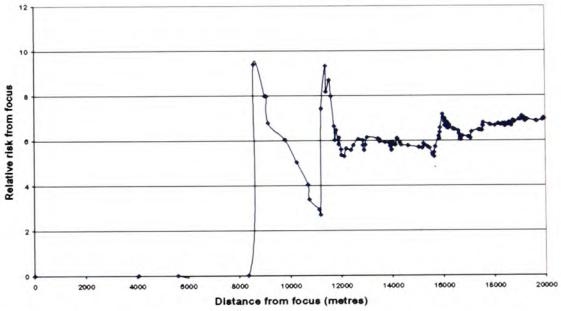


Dataset 2

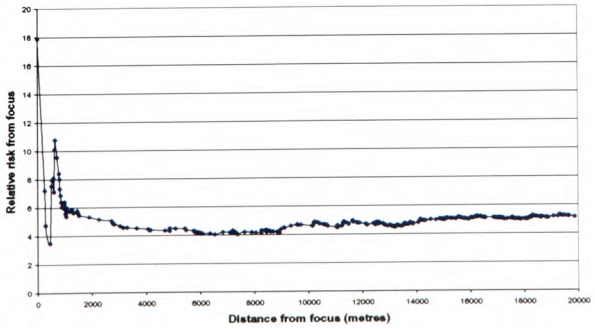
Focus: 414500, 396000



Focus: 461800, 495700

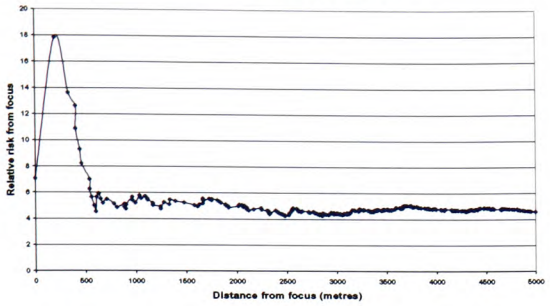


Focus: 436950, 493540

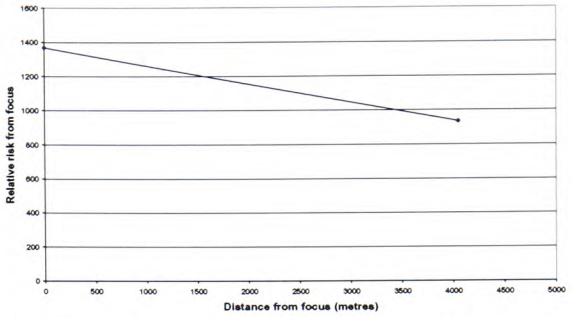


Dataset 03

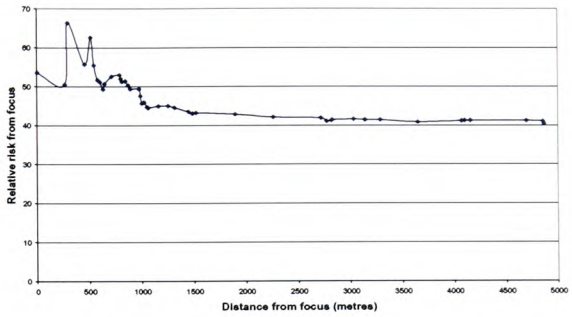
Focus: 414500, 396000



Focus: 461800, 495700

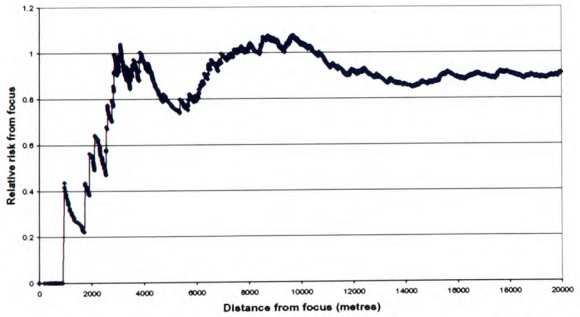


Focus: 436950, 493540

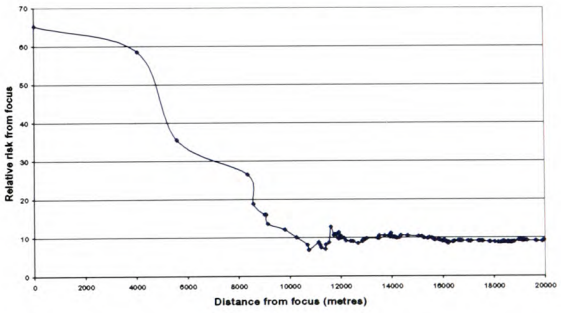


Dataset 4

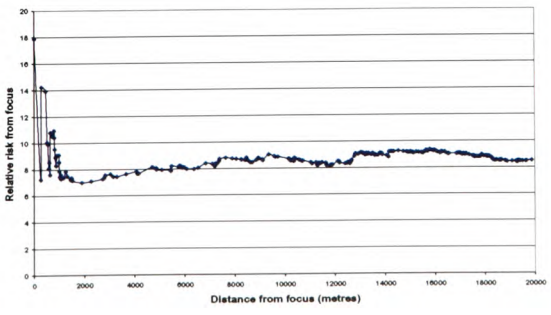
Focus: 414500, 396000



Focus: 461800, 495700



Focus: 436950, 493540



6.4. Appendix D: WinBUGS model used for spatial analysis of survival

```
model;
{
   for( i in 1 : MSOA ) {
       O[i] ~ dnorm(mu[i],tau)
            mu[i] <- E[i] +alpha* b[i]
       b[i] ~ dnorm( b.bar[i], Nneighs[i])
            b.bar[i]    <-  mean(b.neigh[off[i] + 1:off[i+1]])
            SMR[i]      <- (100*mu[i])/E[i]
            Nneighs[i] <-  off[i+1] - off[i]
          }

for(i in 1:neighbours){
            b.neigh[i] <-b[map[i]]
}
            tau    ~ dgamma(0.001, 0.001)
            alpha ~ dgamma(0.001, 0.001)
}
list(MSOA = 413,  neighbours = 2138,

O  =  c(enter all observed survival figures here from MSOA 1
       to MSOA 413),
E  =  c(enter all expected survival figures here from MSOA 1
       to MSOA 413),
map  =  c(enter neighbours for each MSOA in turn from MSOA 1
         to MSOA 413),
off  =  c(0, enter cumulative number of neighbours for each
         MSOA in Wales from MSOA 1 to MSOA 413 (the last
         figure should be 2138)
),
list (initial values)
```

Appendix 7.4: Bayesian model used in WinBUGS.